

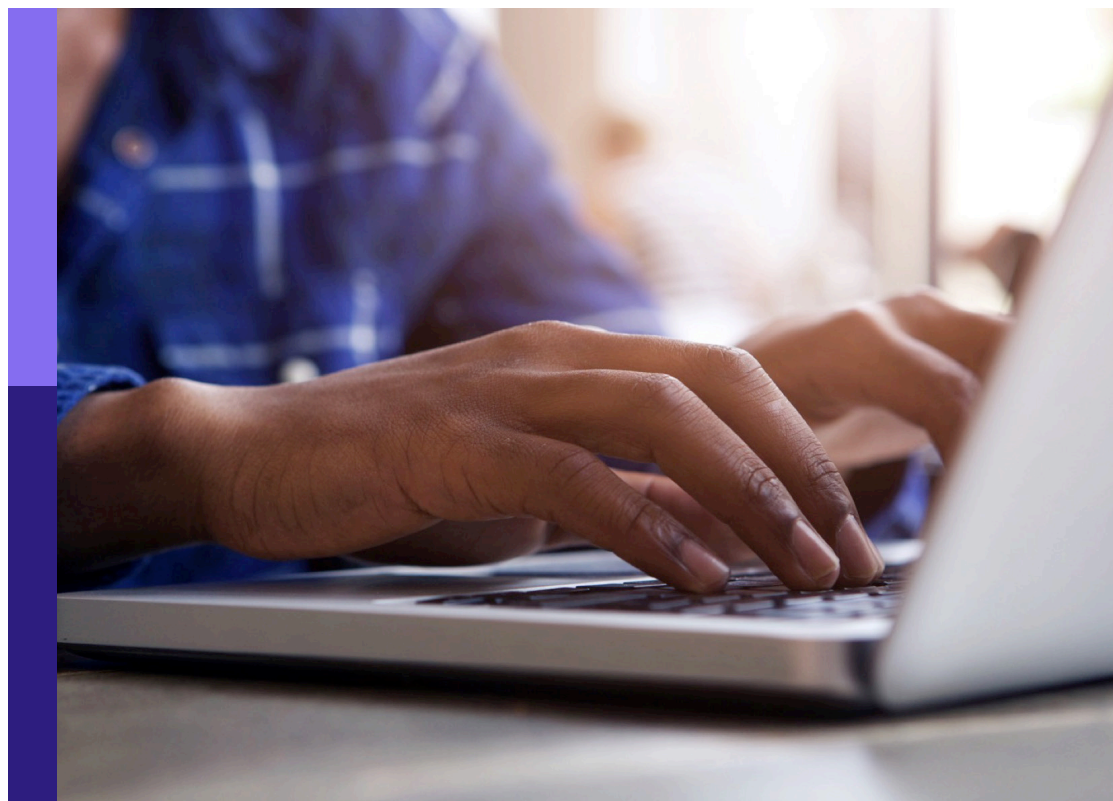
# 2022 Computer science – editor's pick

**Edited by**

Kaleem Siddiqi, Roberto Therón, Kostas Karpouzis,  
Sven Schewe, Nicola Zannone, Marcello Pelillo,  
Kristof Van Laerhoven and Paul Lukowicz

**Published in**

Frontiers in Computer Science



## FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714  
ISBN 978-2-83252-005-5  
DOI 10.3389/978-2-83252-005-5

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: [frontiersin.org/about/contact](https://frontiersin.org/about/contact)

# 2022 Computer science – editor's pick

## Topic editors

Kaleem Siddiqi — McGill University, Canada

Roberto Therón — University of Salamanca, Spain

Kostas Karpouzis — Panteion University, Greece

Sven Schewe — University of Liverpool, United Kingdom

Nicola Zannone — Eindhoven University of Technology, Netherlands

Marcello Pelillo — Ca' Foscari University of Venice, Italy

Kristof Van Laerhoven — University of Siegen, Germany

Paul Lukowicz — University of Kaiserslautern, Germany

## Citation

Siddiqi, K., Therón, R., Karpouzis, K., Schewe, S., Zannone, N., Pelillo, M., Van Laerhoven, K., Lukowicz, P., eds. (2023). *2022 Computer science – editor's pick*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83252-005-5

# Table of contents

04	<b>Eye Movement and Pupil Measures: A Review</b> Bhanuka Mahanama, Yasith Jayawardana, Sundararaman Rengarajan, Gavindya Jayawardena, Leanne Chukoskie, Joseph Snider and Sampath Jayarathna
26	<b>Ten Questions for a Theory of Vision</b> Marco Gori
42	<b>Outer Product-Based Fusion of Smartwatch Sensor Data for Human Activity Recognition</b> Adria Mallol-Ragolta, Anastasia Semertzidou, Maria Pateraki and Björn Schuller
52	<b>Uniform Polylogarithmic Space Completeness</b> Flavio Ferrarotti, Senén González, Klaus-Dieter Schewe and José María Turull-Torres
58	<b>Improving Mobile Device Security by Embodying and Co-adapting a Behavioral Biometric Interface</b> Avinash Jairam, Tzipora Halevi and Theodore Raphan
71	<b>Machine Learning Solutions Applied to Amyotrophic Lateral Sclerosis Prognosis: A Review</b> Fabiano Papaiz, Mario Emílio Teixeira Dourado Jr., Ricardo Alexsandro de Medeiros Valentim, Antonio Higor Freire de Moraes and Joel Perdiz Arrais
83	<b>Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques</b> B. Shamreen Ahamed, Meenakshi Sumeet Arya and Auxilia Osvin Nancy V
88	<b>Micro-HBI: Human-Biology Interaction With Living Cells, Viruses, and Molecules</b> Seung Ah Lee and Ingmar H. Riedel-Kruse
96	<b>Smaller progress measures and separating automata for parity games</b> Daniele Dell'Erba and Sven Schewe
114	<b>Skeletons, Object Shape, Statistics</b> Stephen M. Pizer, J. S. Marron, James N. Damon, Jared Vicory, Akash Krishna, Zhiyuan Liu and Mohsen Taheri
131	<b>An algebra for local histograms</b> Jon Sporring and Sune Darkner
140	<b>The role of gender in the International Conference on Pervasive Computing and Communications</b> Ella Peltonen





# Eye Movement and Pupil Measures: A Review

**Bhanuka Mahanama<sup>1</sup>, Yasith Jayawardana<sup>1</sup>, Sundararaman Rengarajan<sup>2</sup>,  
Gavindya Jayawardena<sup>1</sup>, Leanne Chukoskie<sup>2,3</sup>, Joseph Snider<sup>4</sup> and Sampath Jayarathna<sup>1\*</sup>**

<sup>1</sup>Computer Science, Old Dominion University, Norfolk, VA, United States, <sup>2</sup>Physical Therapy, Movement & Rehabilitation Sciences, Northeastern University, Boston, MA, United States, <sup>3</sup>Art + Design, Northeastern University, Boston, MA, United States, <sup>4</sup>Institute for Neural Computation, UC San Diego, San Diego, CA, United States

Our subjective visual experiences involve complex interaction between our eyes, our brain, and the surrounding world. It gives us the sense of sight, color, stereopsis, distance, pattern recognition, motor coordination, and more. The increasing ubiquity of gaze-aware technology brings with it the ability to track gaze and pupil measures with varying degrees of fidelity. With this in mind, a review that considers the various gaze measures becomes increasingly relevant, especially considering our ability to make sense of these signals given different spatio-temporal sampling capacities. In this paper, we selectively review prior work on eye movements and pupil measures. We first describe the main oculomotor events studied in the literature, and their characteristics exploited by different measures. Next, we review various eye movement and pupil measures from prior literature. Finally, we discuss our observations based on applications of these measures, the benefits and practical challenges involving these measures, and our recommendations on future eye-tracking research directions.

**Keywords:** eye tracking, pupillometry, visual perception, cognition, attention

## OPEN ACCESS

### Edited by:

Kostas Karpouzis,  
Panteion University, Greece

### Reviewed by:

Ilias Maglogiannis,  
University of Piraeus, Greece  
Michael J Proulx,  
University of Bath, United Kingdom

### \*Correspondence:

Sampath Jayarathna  
sampath@cs.odu.edu

### Specialty section:

This article was submitted to  
Human-Media Interaction,  
a section of the journal  
Frontiers in Computer Science

**Received:** 30 June 2021

**Accepted:** 23 November 2021

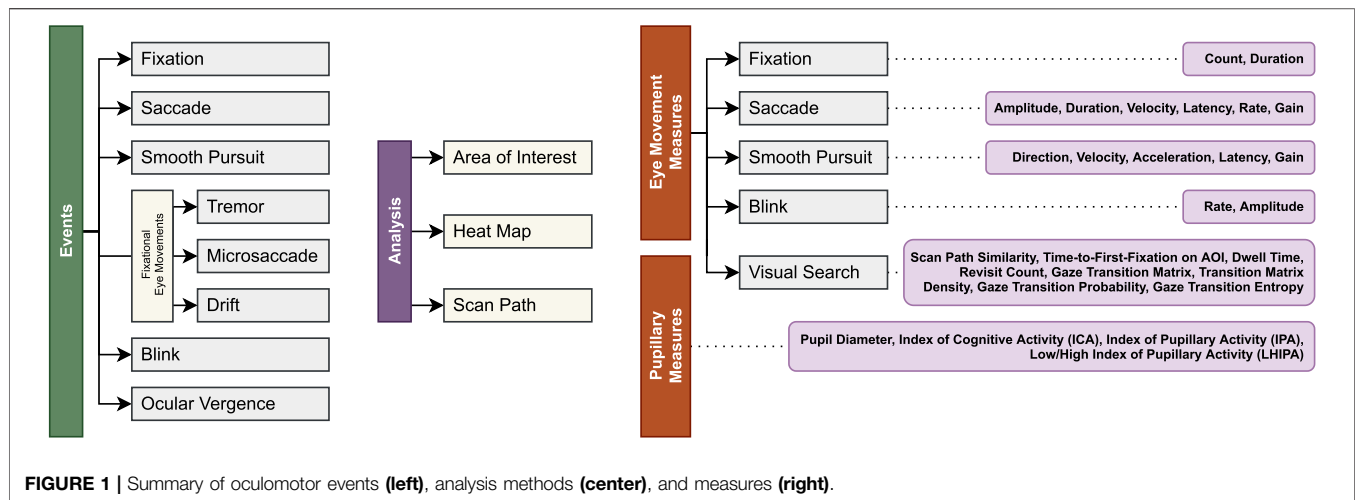
**Published:** 11 January 2022

### Citation:

Mahanama B, Jayawardana Y,  
Rengarajan S, Jayawardena G,  
Chukoskie L, Snider J and  
Jayarathna S (2022) Eye Movement  
and Pupil Measures: A Review.  
Front. Comput. Sci. 3:733531.  
doi: 10.3389/fcomp.2021.733531

## 1 INTRODUCTION

The five primary senses provide humans with a rich perceptual experience of the world, with vision as the dominant sense. Early studies of visual perception (Dodge, 1900; Buswell, 1935; Yarbus, 1967) and its physiological underpinnings (Hubel and Wiesel, 1979; Hubel, 1995), have provided a foundation for subtler and more sophisticated studies of the visual system and its dynamic interaction with the environment via the oculomotor system. The oculomotor system both maintains visual stability and controls gaze-orienting movements (Goldberg et al., 1991; Land and Furneaux, 1997). It is comprised of the *efferent limb* of the visual system and the *vestibular system* (Wade and Jones, 1997). The efferent limb is responsible for maintaining eye position and executing eye movements. The vestibular system, on the other hand, is responsible for providing our brain with information about motion, head position, and spatial orientation, which, in turn, facilitates motor functions, such as balance, stability during movement, and posture (Goldberg and Fernandez, 1984; Wade and Jones, 1997; Day and Fitzpatrick, 2005). The sense of hearing or touch also affects eye movements (Eberhard et al., 1995; Maier and Groh, 2009). There are five distinct types of eye movement, two gaze-stabilizing movements: vestibulo-ocular (VOR), opto-kinetic nystagmus (OKN); and three gaze-orienting movements: saccadic, smooth pursuit, and vergence (Duchowski, 2017; Hejtmancik et al., 2017). For the purposes of this review, we will focus on gaze-orienting eye movements that place the high-resolution fovea on selected objects of interest.



The existence of the fovea, a specialized high-acuity region of the central retina approximately 1–2 mm in diameter (Dodge, 1903), provides exceptionally detailed input in a small region of the visual field (Koster, 1895), approximately the size of a quarter held at arm's length (Pumphrey, 1948; Hejtmancik et al., 2017). The role of gaze-orienting movements are to direct the fovea toward objects of interest. Our subjective perception of a stable world with uniform clarity is a marvel resulting from our visual and oculomotor systems working together seamlessly, allowing us to engage with a complex and dynamic environment.

Recent advancements in computing such as computer vision (Krafka et al., 2016; Lee et al., 2020) and image processing (Pan et al., 2017; Mahanama et al., 2020; Ansari et al., 2021) have led to the development of computing hardware and software that can extract these oculomotor events and measurable properties. This include eye tracking devices range from commodity hardware (Mania et al., 2021) capable of extracting few measures to reserach-grade eye trackers combined with sophisticated software capable of extracting various advance measures. As a result, eye movement and pupillometry have the potential for wide adoption for both in applications and research. There is a need for an aggregate body of knowledge on eye movement and pupillometry measures to provide, a.) a taxonomy of measures linking various oculomotor events, and b.) a quick reference guide for eye tracking and pupillometry measures. For application oriented literature of the eye tracking, interested reader is referred to (Duchowski, 2002) for a breadth-first survey of eye tracking applications.

In this paper, we review a selection of relevant prior research on gaze-orienting eye movements, the periods of visual stability between these movements, and pupil measures to address the aforementioned issue (see **Figure 1**). First, we describe the main oculomotor events and their measurable properties, and introduce common eye movement analysis methods. Next, we review various eye movement and pupil measures. Next, we discuss the applications of aforementioned measures in domains including, but not limited to, neuroscience, human-computer interaction, and psychology, and analyze their strengths and weaknesses. The paper concludes with a

discussion on applications, recent developments, limitations, and practical challenges involving these measures, and our recommendations on future eye-tracking research directions.

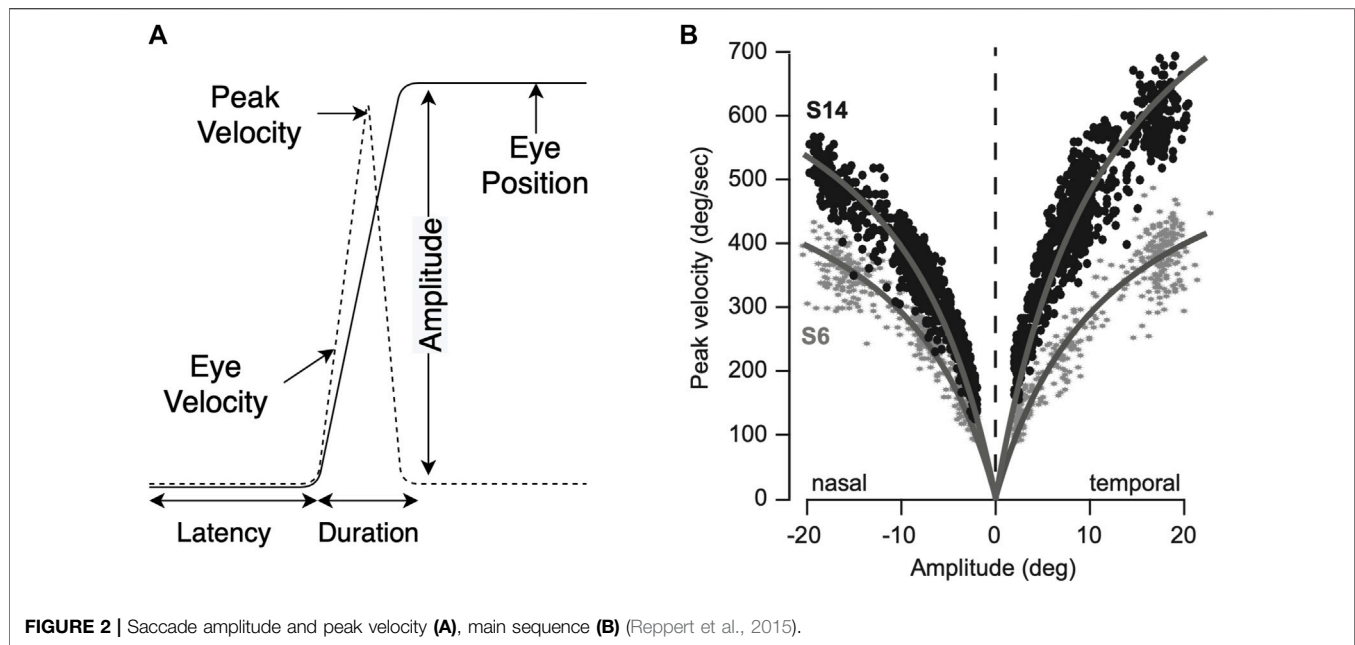
## 2 OCULOMOTOR EVENTS

This section describes oculomotor events that function as the basis for several eye movement and pupil measures. These events are: 1) fixations and saccades, 2) smooth pursuit, 3) fixational eye movements (tremors, microsaccades, drifts), 4) blinks, and 5) ocular vergence.

### 2.1 Fixations and Saccadic Eye Movements

Eye movement information can be interpreted as a sequence of *fixations* and *saccades*. A fixation is a period where our visual gaze remains at a particular location. A saccade, on the other hand, is a rapid eye movement between two consecutive fixations. Typical humans perform 3–5 saccades per second, but this rate varies with current perceptual and cognitive demands (Fischer and Weber, 1993). Fixations and saccades are the primary means of interacting with and perceiving the visual world. During a fixation, our visual perceptual processes unfold. Saccades guide our fovea to selected regions of the visual field. We are effectively blind during saccades (Burr et al., 1994), which allows our gaze to remain relatively stable during saccadic reorientation. Saccadic eye movements are brief, and have a reliable amplitude-velocity relationship (see **Figure 2**) known as the main sequence (Bahill et al., 1975; Termsarasab et al., 2015). It shows that saccade velocity and saccade amplitude follow a linear relationship, up to 15°–20°. This relationship, however, varies with age and also in certain disorders (Choi et al., 2014; Reppert et al., 2015).

Saccades are inhibited during engaged visual attention on stationary stimuli, and as a result, a (nearly) steady central fixation is obtained (Fischer and Breitmeyer, 1987). Previous studies (Schiller et al., 1980; Wang et al., 2015) have shown that saccade preparation processes can be analyzed via pupil size. In particular, Schiller et al. (1980) showed that distinct neural preparatory signals in Superior Colliculus (SC) and Frontal



Eye Field (FEF) are vital for saccade preparation, and Wang et al. (2015) showed that the SC is associated with the pupil control circuit. Cortical processing is associated with saccade latency, with shorter latency indicating advanced motor preparation (Connolly et al., 2005). Reiterating this point, Jainta et al. (2011) showed a negative correlation between saccade latency and pupil size prior to a saccade. Thus, the analysis of measures such as pupil diameter during fixations, fixation duration, saccade rate, saccade accuracy, and saccade latency, provide important cumulative clues to understanding the underlying deployment of visual attention.

### 2.1.1 Identifying Fixations and Saccades

There exists several eye tracking technologies (Young and Sheena, 1975) that measure ocular features over time and transform them into a stream of gaze positions. These streams can be analyzed in different ways to identify periods of fixation and saccades. Salvucci and Goldberg (2000) describe five algorithms for such identification: Velocity Threshold Identification (I-VT), Hidden Markov Model Identification (I-HMM), Dispersion Threshold Identification (I-DT), Minimum Spanning Tree Identification (I-MST), and Area-of-Interest Identification (I-AOI). I-VT and I-HMM are velocity-based algorithms. In I-VT, consecutive points are identified as fixations or saccades, based on their point-to-point velocities (Findlay et al., 1995). I-HMM, on the other hand, uses a two-state Hidden Markov Model with hidden states representing the velocity distributions of saccade and fixation points. Compared to fixed-threshold methods like I-VT, I-HMM performs a more robust identification (Salvucci and Goldberg, 2000), since it employs a probabilistic model rather than a fixed velocity threshold, which allows more freedom in identifying points. I-DT and I-MST are dispersion-based algorithms that use a moving window to calculate the dispersion of points. Based on whether the dispersion is above

or below the threshold, points are classified as fixations or saccades (Widdel, 1984). In I-MST, a minimum-spanning tree is constructed from gaze points. Edges with lengths exceeding a predefined ratio are labeled as saccades and clusters of points connected by saccades are labeled as fixations (Salvucci and Goldberg, 2000). I-AOI is an area-based algorithm that only identifies fixations within specified target areas (AOIs) (Salvucci and Goldberg, 2000). If a point falls within a target area, it is labeled as a fixation point, and if not, it is labeled as a saccade point. Consecutive fixation points are then grouped together, and groups that do not span a minimum duration are re-labeled as saccade points. A systematic evaluation of the performance of these algorithms are available at Komogortsev et al. (2010).

### 2.2 Smooth Pursuit Eye Movements

The smooth pursuit system is a different gaze-orienting movement that is deployed to keep a moving object in foveal vision (Carl and Gellman, 1987; Barnes, 2008). Smooth pursuit eye movements are generally made when tracking an object moving in the visual environment (Carl and Gellman, 1987; Barnes and Asselman, 1991). A typical smooth pursuit movement is usually initiated by a saccadic eye movement to orient to the tracked object. The pursuit system subsequently matches the eye velocity to target velocity (Robinson, 1965; Barnes and Asselman, 1991). This smooth movement is punctuated by additional saccadic movements that eliminate retinal error between the current gaze position and target. The smooth pursuit system has a functional architecture very similar to that of the saccadic system (Lisberger et al., 1987); however, smooth pursuit has a lower latency (100–125 ms) than saccades (200–250 ms) (Meyer et al., 1985; Krauzlis, 2004). Due to the underlying similarity between saccades and smooth pursuits, metrics used to characterize saccades could also be used to characterize smooth pursuit behavior (Lisberger et al., 1987).

According to a classic smooth pursuit behavior model Robinson et al. (1986), there are three aspects of pursuit to characterize: onset, offset and motor learning. Pursuit onset is the response time of the pursuit system to a target which moves for a certain period of time. Since it occurs after the target starts to move, this time delay reflects the response of the pursuit system to the target motion while the eyes were still (Jiang, 1996). Pursuit offset is the response time of the pursuit system to turn off as a response when target stops its motion. When the target is seen to stop, the pursuit system is turned off and replaced by fixation (Robinson et al., 1986). Motor plasticity or motor learning, as a gradual process that makes small, adaptive steps in a consistent direction, was also incorporated and simulated in this model.

## 2.3 Fixational Eye Movements

The process of visual exploration is characterized by alternating fixations and saccades (Salvucci and Goldberg, 2000). However, a fixation does not imply a stationary eye; our eyes are continually in (involuntary) motion, even during periods of fixation (Adler and Fliegelman, 1934; Ratliff and Riggs, 1950; Ditchburn and Ginsborg, 1953). These fixational eye movements fall into one of three classes: 1) *tremors*, 2) *micro-saccades*, and 3) *drifts* (Martinez-Conde et al., 2004; Rucci and Poletti, 2015).

### 2.3.1 Tremor

A tremor (or *ocular micro-tremor*, or *physiological nystagmus*) is an aperiodic, wavelike eye movement with a high-frequency ( $\sim 90$  Hz) (Carpenter, 1988) and low-amplitude ( $\sim$  the diameter of a foveal cone) (Riggs et al., 1953). Due to this nature, tremors fall within the range of recording noise, making it challenging to record them accurately (Carpenter, 1988). Tremors allow the retaining of visual acuity during prolonged fixations (Riggs and Ratliff, 1951; Riggs et al., 1953). For instance, Riggs et al. (1953) showed that when tremors are bypassed artificially, the visual acuity diminishes over time.

### 2.3.2 Microsaccade

A microsaccade (or *flicker*, or *flicker*, or *fixational saccade*) is a small, fast, jerk-like eye movement that occurs during voluntary fixation (Ditchburn and Ginsborg, 1953; Martinez-Conde et al., 2004). They occur at a typical rate of 1–3 Hz, shifting the line of sight abruptly by a small amount (Ditchburn and Ginsborg, 1953). The average size of a microsaccade is about  $6'$  arc (i.e., the size of a thumb-tack head, held 2.5 m away from the eye) (Steinman et al., 1973). The dynamics of microsaccades vary with stimuli and viewing task. For instance, the difficulty of a task can be discerned by the number of microsaccades that occurred (Otero-Millan et al., 2008), and their magnitude (Krejtz et al., 2018). Microsaccades also have comparable spatio-temporal properties as saccades (Zuber et al., 1965; Otero-Millan et al., 2008). For instance, microsaccades lie on the saccadic main sequence (Zuber et al., 1965). The refractory periods between saccades and microsaccades are also equivalent (Otero-Millan et al., 2008). Moreover, microsaccades as small as  $9'$  generate a field potential over the occipital cortex and the mid-central scalp sites 100–140 ms after movement onset, which resembles the visual lambda response evoked by saccades (Dimigen et al., 2009).

It is increasingly accepted that microsaccades play an important role in modulating attentional and perceptual processes (Hafed et al., 2015).

### 2.3.3 Drift

A drift (or *slow drift*) is a low-frequency eye movement that occurs during the intervals between microsaccades and saccades (Steinman et al., 1973). During a drift, the retinal image of a fixated object moves across photoreceptors (Ratliff and Riggs, 1950). Drifts have a compensatory role in maintaining accurate visual fixation; they occur either in the absence of microsaccades, or when the compensation by microsaccades is inadequate (Ratliff and Riggs, 1950). The average size of a drift is about a  $6'$  arc, with an average velocity of about a  $1'$  arc/sec (Ratliff and Riggs, 1950; Ditchburn and Ginsborg, 1953; Cornsweet, 1956; Nachmias, 1961).

## 2.4 Blinks

A blink is essentially the closing and reopening of the eyelids. Blinks are primitive, yet widely used, in eye tracking measures. When the blink originates from a voluntary action, the blink becomes a voluntary blink or a wink (Blount, 1927). In the case of non-voluntary blinks, they are of two types: spontaneous blinks and reflexive blinks. For reflexive blinks, external stimuli evoke reflexive blinks as a form of protection, while any involuntary blink not belonging to any of these categories is a spontaneous blink (Valls-Sole, 2019). The winks or voluntary blinks are not commonly adopted as a metric despite the usage as a form of interaction (Noronha et al., 2017). In contrast, involuntary blinks indicate the state of an individual (Stern et al., 1984) or a reflex action to a stimulus (Valls-Sole, 2019). Between involuntary blinks, spontaneous blinks are the most common type of blink used as a metric due to their correlation with one's internal state (Shin et al., 2015; Maffei and Angrilli, 2019).

## 2.5 Ocular Vergence

Up until this point, the movements described are all referred to as conjugate or “yoked” eye movements, meaning that the eyes move in the same direction to fixate an object. Fortunately, we can choose to fixate on objects in different depth planes, during which binocular vision is maintained by opposite movements of the two eyes. These simultaneously directly opposing movements of the eyes result in Ocular vergence (Holmqvist et al., 2011). These vergence movements can occur in either direction, resulting in convergence or divergence. Far-to-near focus triggers convergent movements and near-to-far focus triggers divergent movements. The ubiquitous use of screen-based eye tracking results in more literature related to conjugate eye movements in a single depth plane.

## 3 EYE MOVEMENT ANALYSIS

Eye movements are a result of complex cognitive processes, involving at the very least, target selection, movement planning, and execution. Analysis of eye movements (see **Table 1** for a list of eye movement measures) can reveal objective and quantifiable information about the quality,

**TABLE 1 |** Summary of related work on eye movement and pupil measures.

Fixation	Related work
Count	Buswell (1935), Yarbus (1967), Schoonahd et al. (1973), Brutton and Janssen (1979), Megaw and Richardson (1979), Megaw (1979), Goldberg and Kotval (1999), Coeckelbergh et al. (2002), Jacob and Karn (2003), Ares et al. (2013)
Duration	Young and Sheena (1975), Rayner (1978), Rayner (1979), Salthouse and Ellis (1980), Karsh and Breitenbach (1983), Goldberg and Kotval (1999), Velichkovsky et al. (2000), Pavlović and Jensen (2009), Staub and Benatar (2013), Ares et al. (2013), Reingold et al. (2012), Menon et al. (2016), Costa et al. (2018), Henderson et al. (2018), Velichkovsky et al. (2019)
SACCADE	Related Work
Amplitude	Bahill et al. (1975), Ceder (1977), Rayner (1978), Megaw and Richardson (1979), May et al. (1990), Zelinsky and Sheinberg (1997), Phillips and Edelman (2008), Rayner et al. (2012), Anson et al. (2016), Buonocore et al. (2016), Buonocore et al. (2017), Le Meur et al. (2017), Mostofi et al. (2020)
Direction	Takeda and Funahashi (2002), Killian et al. (2015), Walker et al. (2006), Foulsham et al. (2008), Ponsoda et al. (1995), Gbadamosi and Zangemeister (2001), Yu et al. (2016), Mulder et al. (2020), Anderson et al. (2020)
Velocity	Becker and Fuchs (1969), Lehtinen et al. (1979), Griffiths et al. (1984), Abel and Hertle (1988), Galley (1993), McGregor and Stern (1996), Castello et al. (1998), Russo et al. (2003), Xu-Wilson et al. (2009), Boxer et al. (2012), Buonocore et al. (2016), Buonocore et al. (2017), Mostofi et al. (2020)
Latency	McKee et al. (2016), Warren et al. (2013), Michell et al. (2006), McSorley et al. (2012), Knox and Wolohan (2014), Anson et al. (2016), Lai et al. (2019)
Rate	Ohtani (1971), Van Orden et al. (2000), Nakayama et al. (2002), O'Driscoll and Callahan (2008), van Tricht et al. (2010)
Gain	Coëffé and O'regan (1987), Ettinger et al. (2002), Crevits et al. (2003), Lisi et al. (2019)
SMOOTH PURSUIT	Related Work
Direction	Collewijn and Tamminga (1984), Rottach et al. (1996)
Velocity	Barmack (1970), Young (1971), Bahill and Laritz. (1984), Meyer et al. (1985), De Brouwer et al. (2002)
Acceleration	Kao and Morrow (1994), Ladda et al. (2007)
Latency	Braun et al. (2006), Burke and Barnes (2006), de Hemptinne et al. (2006), Spering and Gegenfurtner (2007)
Retinal Position Error	de Brouwer et al. (2001)
Gain	Robinson (1965), Zackon and Sharpe (1987), Rottach et al. (1996), Churchland and Lisberger (2002), O'Driscoll and Callahan (2008)
BLINK	Related Work
Rate	Newhall (1932), Doughty (2001), Doughty (2002), Doughty and Naase (2006), Oh et al. (2012), Shin et al. (2015), Jongkees and Colzato (2016), Doughty (2019), Maffei and Angrilli (2019), Ranti et al. (2020)
Amplitude	Stevenson et al. (1986), Riggs et al. (1987), Morris and Miller (1996), Galley et al. (2004), Cardona et al. (2011), Chu et al. (2014)
VISUAL SEARCH	Related Work
Scan Path Similarity	Jarodzka et al. (2010), De Bruin et al. (2013)
Time-to-First-Fixation on AOI	Krupinski (1996), Ellis et al. (1998), Jacob and Karn (2003), Bojko (2006), Venjakob et al. (2012), Donovan and Litchfield (2013)
Dwell Time	Tullis and Albert (2013), Mohanty and Sussman (2013), Hüsser and Wirth (2014), Ceravolo et al. (2019)
Revisit Count	Guo et al. (2016), Meghanathan et al. (2019), Mello-Thoms et al. (2005), Motoki et al. (2021)
Gaze Transition Matrix	Ponsoda et al. (1995), Bednarik et al. (2005)
Transition Matrix Density	Goldberg and Kotval (1999)
Gaze Transition Probability	Vandenberg et al. (2013), Jayawardena et al. (2020)
Gaze Transition Entropy	Krejtz et al. (2014), Krejtz et al. (2015), Shiferaw et al. (2019)
VERGENCE	Related Work
Ocular Vergence	Daugherty et al. (2010), Essig et al. (2004), Wang et al. (2012), Mlot et al. (2016)
PUPIL	Related Work
Diameter	Gray et al. (1993), Joshi et al. (2016), Rubaltelli et al. (2016)
ICA	Marshall (2000), Marshall (2002), Marshall (2007), Abel and Hertle (1988), Bartels and Marshall (2012), Demberg (2013), Demberg et al. (2013), Korbach et al. (2017), Korbach et al. (2018), Rerhaye et al. (2018)
IPA	Duchowski et al. (2018), Krejtz et al. (2020), Fehring (2020), Fehring (2021)
LHIPA	Duchowski et al. (2020), Krejtz et al. (2020)

predictability, and consistency of these covert processes (Van der Stigchel et al., 2007). Holmqvist et al. (2011) and Young and Sheena (1975) discuss several eye movement measures, eye

movement measurement techniques, and key considerations for eye movement research. In this section, we introduce several eye movement analysis techniques from the literature.



### 3.1 Area of Interest Analysis

AOI analysis is a technique to analyze eye movements by assigning them to specific *areas* (or regions) of the visual scene (Holmqvist et al., 2011; Hessels et al., 2016). In contrast to obtaining eye movement measures across the entire scene, AOI analysis provides semantically localized eye movement measures that are particularly useful for attention-based research (e.g., User Interaction research, Marketing research, and Psychology research) (Hessels et al., 2016). In AOI analysis, defining the shape and bounds of an AOI can be difficult (Hessels et al., 2016). Ideally, each AOI should be defined with the same shape and bounds as the actual object. However, due to practical limitations, such as the difficulty of defining arbitrarily shaped (and sized) AOIs in eye-tracking software, AOIs are most commonly defined using simple shapes (rectangles, ellipses, etc.) (Holmqvist et al., 2011). Recent advancements in computer vision have given rise to models that automatically and reliably identify real world objects in visual scenes Ren et al. (2015), Redmon et al. (2016). This makes it possible to identify AOIs in real world images almost as readily as with pre-defined stimuli presented on a computer screen Jayawardena and Jayarathna (2021), Zhang et al. (2018).

### 3.2 Heat Map Analysis

Heat map analysis is a technique for analyzing the spatial distribution of eye movements across the visual scene. This technique can be used to analyze eye movements of individual participants, as well as aggregated eye movements of multiple participants. In general, heat maps are represented using Gaussian Mixture Models (GMMs) that indicate the frequency (or probability) of fixation localization. Heat map-based metrics generally involve a measure of overlap between two GMMs, indicating similarity of fixated regions of an image. In heat map analysis, the order of visitation is not captured (Grindinger et al., 2010); rather, it analyzes the spatial distribution of fixations. When visualizing heat maps, a color-coded, spatial distribution of fixations is overlaid on the stimuli that participants looked at. The color represents the quantity of fixations at each point on the heat map. Heat map analysis is particularly useful when analyzing the areas of the stimuli that participants paid more (or less) visual attention to, for example in driving research comparing different groups and conditions (Snider et al., 2021).

### 3.3 Scan Path Analysis

A *scan path* is a sequence of fixations and saccades that describe the pattern of eye movements during a task (Salvucci and Goldberg, 2000). Scan paths could appear quite complex, with frequent revisits and overlapping saccades. In general, scan paths are visualized as a sequence of connected nodes (a fixation centroid) and edges (a saccade between two successive fixations) displayed over the visual image of the scene (Hemingshous and Duchowski, 2006; Goldberg and Helfman, 2010). Here, the diameter of each node is proportional to the fixation duration (Goldberg and Helfman, 2010). Scan path analyses have been widely used to model the dynamics of eye movement during visual search (Walker-Smith et al., 1977; Horley et al., 2003). It has also been used in areas like

biometric identification (Holland and Komogortsev, 2011). In cognitive neuroscience, Parkhurst et al. (2002) and van Zoest et al. (2004) show that stimulus-driven, bottom-up attention dominates during the early phases of viewing. On the contrary, Nyström and Holmqvist (2008) show that top-down cognitive processes guide fixation selection throughout the course of viewing. They discovered that viewers eventually fixate on meaningful stimuli, regardless of whether that stimuli was obscured or reduced in contrast.

## 4 EYE MOVEMENT MEASURES

In this section, we discuss several metrics relevant to oculomotor behavior (Komogortsev et al., 2013) which are derived from fixations, saccades, smooth pursuit, blinks, vergence, and visual search paradigm.

### 4.1 Fixation Measures

Fixation-based measures are widely used in eye-tracking research. Here, fixations are first identified using algorithms such as I-VT, I-HMM (velocity-based), I-DT, I-MST (dispersion-based), and I-AOI (area-based) (Salvucci and Goldberg, 2000). This information is then used to obtain different fixation measures, as described below.

#### 4.1.1 Count

Fixation count is the number of fixations identified within a given time period. Fixations can be counted either over the entire stimuli or within a single AOI (Holmqvist et al., 2011). Fixation count has been used to determine semantic importance (Buswell, 1935; Yarbush, 1967), the efficiency and difficulty in search (Goldberg and Kotval, 1999; Jacob and Karn, 2003), neurological dysfunctions (Brutten and Janssen, 1979; Coeckelbergh et al., 2002), and the impact of prior experience (Schoonahd et al., 1973; Megaw, 1979; Megaw and Richardson, 1979).

#### 4.1.2 Duration

Fixation duration indicates a time period where the eyes stay still in one position (Salvucci and Goldberg, 2000). In general, fixation durations are around 200–300 ms long, and longer fixations indicate deeper cognitive processing (Rayner, 1978; Salthouse and Ellis, 1980). Also, fixations could last for several seconds (Young and Sheena, 1975; Karsh and Breitenbach, 1983) or be as short as 30–40 ms (Rayner, 1978; Rayner, 1979). Furthermore, the distribution of fixation duration is typically positively skewed, rather than Gaussian (Velichkovsky et al., 2000; Staub and Benatar, 2013). The *average* fixation duration is often used as a baseline to compare with fixation duration data at different levels (Salthouse and Ellis, 1980; Pavlović and Jensen, 2009). By comparing average fixation duration across AOIs, one could distinguish areas that were looked at for longer durations than others. In particular, if certain AOIs were looked at longer than others, then their average fixation duration would be higher (Goldberg and Kotval, 1999; Pavlović and Jensen, 2009).

## 4.2 Saccade Measures

### 4.2.1 Amplitude

The amplitude of a saccade (see **Figure 2**) is the distance travelled by a saccade during an eye movement. It is measured either by visual degrees (angular distance) or pixels, and can be approximated via the Euclidean distance between fixation points (Megaw and Richardson, 1979; Holmqvist et al., 2011). Saccade amplitudes are dependent on the nature of the visual task. For instance, in reading tasks, saccade amplitudes are limited to  $\approx 2^\circ$ , i.e., 7–8 letters in a standard font size (Rayner, 1978). They are further limited when oral reading is involved (Rayner et al., 2012). Furthermore, saccade amplitudes tend to decrease with increasing task difficulty (Zelinsky and Sheinberg, 1997; Phillips and Edelman, 2008), and with increasing cognitive load (May et al., 1990; Ceder, 1977).

### 4.2.2 Direction

The direction (or *orientation*, or *trajectory*) of a saccade or sequence of saccades is another useful descriptive measure. It can be represented as either an absolute value, a relative value, or a discretized value. Absolute saccade direction is calculated using the coordinates of consecutive fixations. Relative saccade direction is calculated using the difference of absolute saccade direction of two consecutive saccades. Discretized saccade directions are obtained by binning the absolute saccade direction into pre-defined angular segments (e.g., compass-like directions). In visual search studies, researchers have used different representations of saccade direction to analyze how visual conditions affect eye movement behavior. For instance, absolute saccade directions were used to analyze the effect of target predictability (Walker et al., 2006) and visual orientation (Foulsham et al., 2008) on eye movements. Similarly, discretized saccade directions were used to compare and contrast the visual search strategies followed by different subjects (Ponsoda et al., 1995; Gbadamosi and Zangemeister, 2001).

### 4.2.3 Velocity

Saccade velocity is calculated by taking the first derivative of time series of gaze position data. Average saccadic velocity is the average of velocities over the duration of a saccade. Peak saccadic velocity is the highest velocity reached during a saccade (Holmqvist et al., 2011). For a particular amplitude, saccade velocity has been found to decrease with tiredness (Becker and Fuchs, 1969; McGregor and Stern, 1996), sleep deprivation (Russo et al., 2003), and conditions such as Alzheimer's (Boxer et al., 2012) and AIDS (Castello et al., 1998). In contrast, saccade velocity has been found to increase with increasing task difficulty (Galley, 1993), increasing intrinsic value of visual information (Xu-Wilson et al., 2009), and increasing task experience (McGregor and Stern, 1996). Many studies on neurological and behavioral effects of drugs and alcohol (Lehtinen et al., 1979; Griffiths et al., 1984; Abel and Hertle, 1988) have used peak saccade velocity as an oculomotor measure.

### 4.2.4 Latency

Saccadic latency measures the duration between the onset of a stimulus and the initiation of the saccade (Andersson et al., 2010).

In practice, the measurement of the saccadic latency is affected by two main factors: The sampling frequency of the setup, and the saccade detection time. The sampling frequency refers to the operational frequency of the eye tracker, where the operational frequency negatively correlates with the introduced error. The saccade detection time is when the device detects a saccade by arriving at the qualifying velocity or the criteria in the saccade detection algorithm. In a study with young and older participants, researchers observed the saccadic latencies to increase significantly in older participants with the decrease in the stimulus size Warren et al. (2013). Further, in comparative studies between healthy subjects and subjects with Parkinson's disease with and without medication, saccadic latency shows potential as a biomarker for the disease (Michell et al., 2006). A similar study with participants having amblyopia has found the interocular difference between saccadic latencies to correlate with the difference in Snellen acuity (McKee et al., 2016).

### 4.2.5 Rate

Saccade rate (or *saccade frequency*) is the number of saccadic eye movements per unit time (Ohtani, 1971). For static stimuli, the saccade rate is similar to the fixation rate. For dynamically moving stimuli, however, the saccade rate is a measure of catch-up saccades generated during smooth pursuit (Holmqvist et al., 2011). The saccade rate decreases with increasing task difficulty (Nakayama et al., 2002) and fatigue level (Van Orden et al., 2000). Moreover, subjects with neurological disorders exhibit higher saccadic rates during smooth pursuit (O'Driscoll and Callahan, 2008; van Tricht et al., 2010).

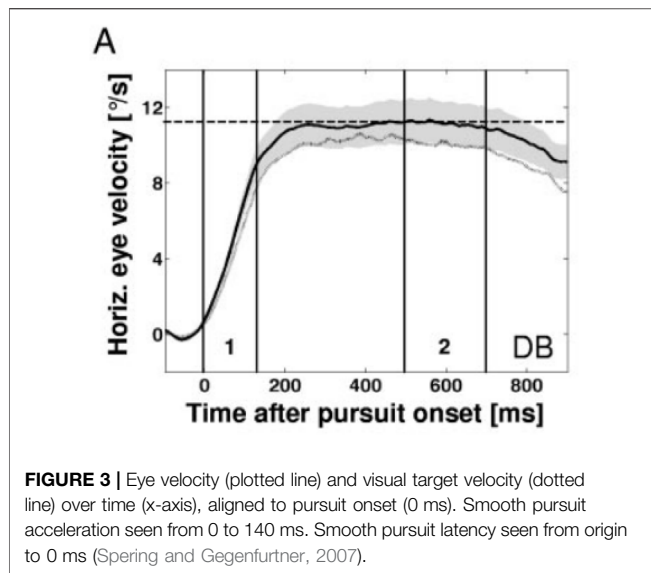
### 4.2.6 Gain

Saccade gain (or *saccade accuracy*) is the ratio between the initial saccade amplitude and the target amplitude (i.e., Euclidean distance between the two stimuli among which that saccade occurred) (Coëffé and O'regan, 1987). This measure indicates how accurately a saccadic movement landed on the target stimuli (Ettinger et al., 2002; Holmqvist et al., 2011). When the gain of a particular saccade is greater than 1.0, that saccade is called an *overshoot*, or *hypermetric*, and when it is less than 1.0, that saccade is called an *undershoot*, or *hypometric*. Saccadic gain is probabilistic at a per-individual level; Lisi et al. (2019) demonstrates that biased saccadic gains are an individualized probabilistic control strategy that adapts to different environmental conditions. Saccade gain is commonly used in neurological studies (Ettinger et al., 2002; Holmqvist et al., 2011). For instance, Crevits et al. (2003) used saccade gain to quantify the effects of severe sleep deprivation.

## 4.3 Smooth Pursuit Measures

### 4.3.1 Direction

Smooth pursuit direction (or *smooth pursuit trajectory*) indicates the direction of smooth pursuit movement as the eyes follow a moving stimulus (Holmqvist et al., 2011). The ability to pursue a moving object varies with its direction of motion. Rottach et al. (1996) showed that smooth pursuit gain is higher during horizontal pursuit than during vertical pursuit. This difference



in gain can be attributed to most real-world objects naturally being in horizontal rather than vertical motion (Collewijn and Tamminga, 1984).

#### 4.3.2 Velocity

Smooth pursuit velocity is first moment of gaze positions during a smooth pursuit. Compared to saccade velocity, the velocity of smooth pursuits is low, typically around 20°/s–40°/s (Young, 1971). However, when participants are specifically trained to follow moving stimuli, or are provided with accelerating stimuli, higher peak smooth pursuit velocities were observed (Meyer et al., 1985). For example, Barmack (1970) observed peak smooth pursuit velocities of 100°/s in typical participants, when provided with accelerating stimuli. However, Bahill and Laritz (1984) observed peak smooth pursuit velocities of 130°/s on trained baseball players. As the velocity of moving stimuli increases, the frequency of catch-up saccades increases to compensate for retinal offset (De Brouwer et al., 2002).

#### 4.3.3 Acceleration

The second moment of the gaze position trace provides the acceleration of a smooth pursuit eye movement. This acceleration is maintained until the eye velocity (smooth pursuit velocity) matches the visual target's velocity (Kao and Morrow, 1994). Examination of acceleration is typically a part of determining smooth pursuit onset (see Figure 3). Smooth pursuit acceleration has been used to analyze how visual cues (Ladda et al., 2007) and prior knowledge of a visual target's trajectory (Kao and Morrow, 1994) impact eye movements. Smooth pursuit acceleration is higher when a visual target's motion was unpredictable.

#### 4.3.4 Latency

Smooth pursuit latency is the delay between when a target object starts to move (i.e., target onset) and when the pursuit begins (i.e., smooth pursuit onset) (Holmqvist et al., 2011). When the

direction and velocity of the target object are not predictable, the smooth pursuit latency varies between 100–200 ms (Burke and Barnes, 2006). In paradigms such as step-ramp (Rashbass, 1961) that allow for anticipation, smooth pursuit latency may drop to 0 ms (or less, if pursuit starts before target motion) when its direction and velocity are predictable (Burke and Barnes, 2006; de Hemptinne et al., 2006). If the luminance of the moving object is the same as the background, the smooth pursuit latency may be prolonged by ~50 ms (Braun et al., 2006). Smooth pursuit latency is also affected by distracting motion (Spering and Gegenfurtner, 2007); latency was increased when a distractor moved parallel to the pursuit direction but was decreased when the distractor moved opposite to the pursuit direction.

#### 4.3.5 Retinal Position Error

Fixations are maintained more accurately on stationary targets rather than moving targets. During smooth pursuit, both the eye and the target are in motion, and lag between their positions is expected (Dell'Osso et al., 1992). This error is known as retinal position error, and is formally defined as the difference between eye and target positions measured during fixations.

#### 4.3.6 Gain

Smooth pursuit gain (or *smooth pursuit accuracy*) is the ratio between smooth pursuit velocity and the target velocity (Zackon and Sharpe, 1987; Holmqvist et al., 2011). Typically, the smooth pursuit gain is lower than 1.0 and tends to fall even lower when the target velocity is high (Zackon and Sharpe, 1987). Moreover, smooth pursuit gain is modulated by on-line gain control (Robinson, 1965; Churchland and Lisberger, 2002). Smooth pursuit gain is decreased during conditions that distract user attention (Březinová and Kendell, 1977). Furthermore, smooth pursuit gain is also higher when tracking horizontal motion, compared to tracking vertical motion Rottach et al. (1996). Smooth pursuit gain has also been used in neurologically research. For instance, O'Driscoll and Callahan (2008) analyzed the smooth pursuit gain of individuals diagnosed with schizophrenia, and observed a low smooth pursuit gain in these populations. The smooth pursuit gain can be quantified by the root mean squared distance between the target point and the gaze position ( $\theta$ ) over the span of the experiment of  $n$  data samples.

$$\theta_{RMSE} = \sqrt{\frac{\theta_1^2 + \theta_2^2 + \dots + \theta_n^2}{n}} \quad (1)$$

### 4.4 Blink Measures

#### 4.4.1 Rate

Blink rate (or *spontaneous blink rate*, or *blink frequency*) is typically measured in blinks per minute. In some studies, the time between blinks (or blink interval) is measured instead (Shin et al., 2015). Early studies (Peterson and Allison, 1931; Newhall, 1932) show that blink rate is subjected to factors such as lighting, time of the day (fatigue), temperature, wind, age, and sex. Moreover, while blinks are predominantly involuntary, they are inhibited during engaged visual attention to minimize any



blink-induced interruption to visual information (Ranti et al., 2020). More recent studies explore the idea of standard spontaneous eye blink rate in a broader aspect; healthy and non-healthy individuals under single (Doughty, 2002; Doughty and Naase, 2006; Ranti et al., 2020) and multiple (Doughty, 2001; Doughty, 2019) experiment conditions. Their results demonstrate the lack of a common value for blink rate, as they are dependent on experimental conditions. Thus, a standard measure for blink rate remains illusive despite the vast body of studies. Many studies on cognition adopt eye blink rate as a metric due its relative ease of detection, and its ability to indicate aspects of one's internal state. In a study on the impact of blink rate on a Stroop task, researchers found that blink rate increases during a Stroop task compared with a baseline task of resting (Oh et al., 2012). Further, a study involving movie watching found the blink rate a reliable biomarker for assessing the concentration level (Shin et al., 2015; Maffei and Angrilli, 2019). Studies in dopamine processes also use blink rate as a marker of dopamine function. A review on studies in cognitive dopamine functions and the relationship with blink rate indicates varying results on blink rate in dopaminergic studies, including primates (Jongkees and Colzato, 2016).

#### 4.4.2 Amplitude

The blink amplitude is the measure of the distance traveled by (the downward distance of upper eyelid) in the event of a blink (Stevenson et al., 1986). The amplitude measures the relative distance of motion to the distance the eyelid travels in a complete blink. This measure can be obtained using a video-based eye tracker (Riggs et al., 1987) or using electrooculography (Morris and Miller, 1996). Blink amplitude is often used in conjunction with research associated with fatigue measurement (Morris and Miller, 1996; Galley et al., 2004) and task difficulty (Cardona et al., 2011; Chu et al., 2014). Cardona et al. (2011) assessed the characteristics of blink behavior during visual tasks requiring prolonged periods of demanding activities. During the experiment, the authors noted a larger percentage of incomplete blinks. Another study involving fatigued pilots in a flight simulator showed that blink amplitude served as a promising predictor for level of fatigue (Morris and Miller, 1996).

### 4.5 Visual Search Measures

Visual search behavior combines instances of saccades, fixations and possibly also smooth pursuit. The methods used to analyze such sequences of behavior are described below.

#### 4.5.1 Scan Path Similarity

Vector and string-based editing approaches have been developed to compute the similarity of scan paths (Jarodzka et al., 2010; De Bruin et al., 2013). In particular, De Bruin et al. (2013) introduced three metrics: 1) *Saccade Length Index (SLI)*, 2) *Saccade Deviation Index (SDI)*—to assist in faster analysis of eye tracking data, and 3) *Benchmark Deviation Vectors (BDV)*—to highlight repetitive path deviation in eye tracking data. The SLI is the sum of the distance of all the saccades during the experiment. The SLI can be

obtained through following equation, where  $s$  is the starting position of the saccade,  $e$  ending position of the saccade, and  $n$  is the total number of saccades.

$$SLI_{Total} = \sum_{i=1}^n \sqrt{(s_x - e_x)^2 + (s_y - e_y)^2} \quad (2)$$

Jarodzka et al. (2010), on the other hand, proposed representing scan paths as geometrical vectors, and simplifying scan paths by clustering consecutive saccades directed at most  $T\phi$  radians apart.

#### 4.5.2 Time-to-First-Fixation on AOI

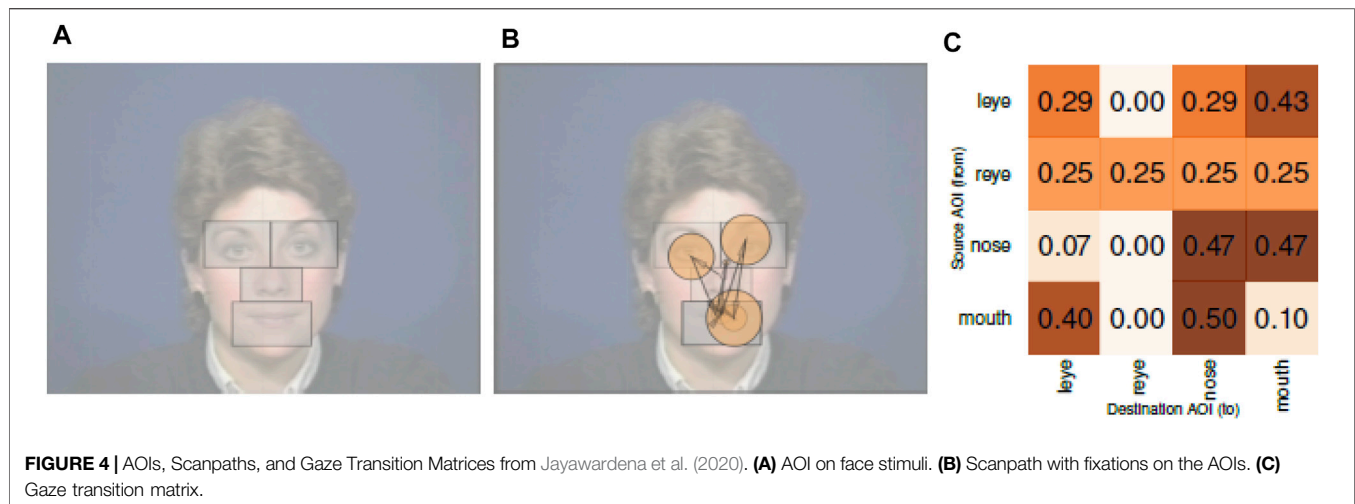
Time-to-First-Fixation on AOI (or *time to first hit*) refers to the time taken from stimulus onset up to the first fixation into a particular AOI (Holmqvist et al., 2011). This measure may be useful for both bottom-up stimulus-driven searches (e.g., a flashy company label) as well as top-down attention driven searches (e.g., when respondents actively decide to focus on certain elements or aspects on a website or picture). This metric is particularly useful for user interface evaluation, as a measure of visual search efficiency (Jacob and Karn, 2003). For instance, it has been used to evaluate the visual search efficiency of web pages (Ellis et al., 1998; Bojko, 2006). It is also influenced by prior knowledge, such as domain expertise. For instance, studies show that when analyzing medical images, expert radiologists exhibit a lower Time to First Fixation on AOIs (lesions, tumors, etc.) than novice radiologists (Krupinski, 1996; Venjakob et al., 2012; Donovan and Litchfield, 2013).

#### 4.5.3 Revisit Count

Revisit (or *re-fixation*, or *recheck*) count indicates how often the gaze was returned to a particular AOI. It can be used to distinguish between the AOIs that were frequently revisited, and the AOIs that were less so. A participant may be drawn back to a particular AOI for different reasons, such as its semantic importance (Guo et al., 2016), to refresh the memory (Meghanathan et al., 2019), and for confirmatory purposes (Mello-Thoms et al., 2005). The emotion perceived through visual stimuli also affect the likelihood of subsequent revisits, and thereby, the revisit count (Motoki et al., 2021); yet this perceived emotion is difficult to interpret purely through revisit count. Revisits are particularly common in social scenes, where observers look back and forth between interacting characters to assess their interaction (Birmingham et al., 2008). Overall, revisit count is indicative of user interest towards an AOI, and can be used to optimize user experiences.

#### 4.5.4 Dwell Time

Dwell time is the interval between one's gaze entering an AOI and subsequently exiting it (Holmqvist et al., 2011). This includes the time spent on all fixations, saccades, and revisits during that visit (Tullis and Albert, 2013). For a typical English reading task, a lower dwell time (e.g., < 100ms) may imply limited information processing, and a higher dwell time (e.g., > 500ms) may imply otherwise (Tullis and Albert, 2013). Dwell times are dependent on factors such as the size and complexity of content within an AOI



(Goldberg and Kotval, 1999), the level of interest towards an AOI (Fisher et al., 2017), situational awareness (Hauland and Duijm, 2002), and task difficulty (Goldberg and Kotval, 1999). In some cases, a higher dwell time could be associated with motivation and top-down attention, as goal-driven respondents may refrain from looking at contextually irrelevant stimuli (Mohanty and Sussman, 2013).

#### 4.5.5 Gaze Transition Matrix

The gaze transition matrix is an adaptation of the transition matrix of Markov models into eye movement analysis. In a Markov model, a *transition matrix* (or *probability matrix*, or *stochastic matrix*, or *substitution matrix*) is a square matrix, where each entry represents the transition probability from one state to another. This concept was first applied for eye movement analysis by Ponsoda et al. (1995) to model the transition of *saccade direction* during visual search. Similarly, Bednarik et al. (2005) used this concept to model the transition of *gaze position* among AOIs and study the correlation between task performance and search pattern. The gaze transition matrix is calculated using the number of transitions from the  $X^{\text{th}}$  AOI to the  $Y^{\text{th}}$  AOI. Based on the gaze transition matrix, several measures have been introduced to quantify different aspects of visual search.

#### 4.5.6 Transition Matrix Density

Goldberg and Kotval (1999) defined *transition matrix density* (i.e., fraction of non-zero entries in the transition matrix) in order to analyze the efficiency of visual search. Here, a lower transition matrix density indicates an efficient and directed search, whereas a dense transition matrix indicates a random search.

#### 4.5.7 Gaze Transition Probability

When comparing gaze transition matrices, Vandenberg et al. (2013) performed an element-wise comparison of transition probabilities by modelling eye movement transitions as a multi-level Hidden Markov Model. Similarly, Jayawardena et al. (2020) utilized gaze transition matrices to analyze the probabilities of transition of gaze between four AOIs.

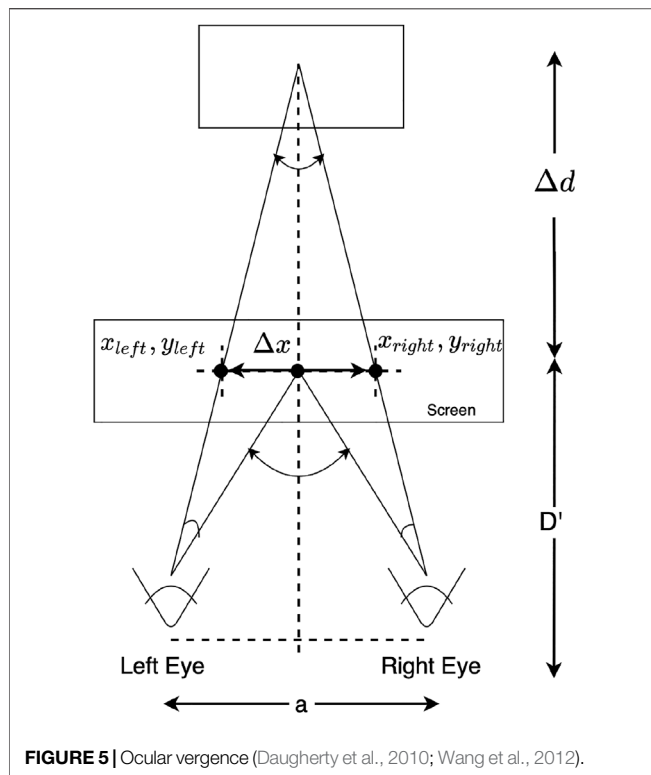
**Figure 4A** shows the AOIs used in this study. **Figure 4B** shows a sample scanpath of a participant, and **Figure 4C** shows its corresponding gaze transition matrix. Here, each cell in the matrix represents the probability of gaze transition from one AOI to another.

#### 4.5.8 Gaze Transition Entropy

Gaze transition entropy is a measure of predictability in AOI transitions and overall distribution of eye movements over stimuli (Krejtz et al., 2014; Krejtz et al., 2015). Compared to transition matrix density, gaze transition entropy is a histogram-based estimation, which, in effect, takes into account the AOI size. The concept of entropy used here is that of information theory; it describes the amount of information required to generate a particular sequence, as a measure of its uncertainty. Krejtz et al. (2014), Krejtz et al. (2015) computes gaze transition entropy by first modelling eye movements between AOIs as a first-order Markov chain, and then obtaining its Shannon's entropy (Shannon, 1948). They obtain two forms of entropy: 1) transition entropy  $H_t$  (calculated for individual subjects' transition matrices) and 2) stationary entropy  $H_s$  (calculated for individual subjects' stationary distributions).

$$H_t = - \sum_i \pi_i \sum_j p_{ij} \log_2 p_{ij} \quad H_s = - \sum_i \pi_i \log_2 \pi_i \quad (3)$$

$H_t$  indicates the predictability of gaze transitions; a high  $H_t$  ( $\forall_{ij} |p_{ij} \rightarrow 0.5$ ) implies low predictability, whereas a low  $H_t$  ( $\forall_{ij} |p_{ij} \rightarrow \{0, 1\}$ ) implies high predictability. It is calculated by normalizing the transition matrix row-wise, replacing zero-sum rows by the uniform transition probability  $1/s$  (Here,  $s$  is the number of AOIs), and obtaining its Shannon's entropy.  $H_s$ , on the other hand, indicates the distribution of visual attention (Krejtz et al., 2014); a high  $H_s$  indicates that visual attention is equally distributed across all AOIs, whereas a low  $H_s$  indicates that visual attention is directed towards certain AOIs. It is estimated via eigen-analysis, and the Markov chain is assumed to be in a steady state where the transition probabilities converge. In one study, Krejtz et al. (2015) showed that participants who viewed artwork with a reportedly high curiosity, yielded a



significantly less  $H_t$  (i.e., more predictable transitions) than others. Moreover, participants who viewed artwork with a reportedly high appreciation, yielded a significantly less  $H_s$  (i.e., more directed visual attention) than others. In another study, Krejtz et al. (2015) showed that participants who reportedly recognized a given artwork, yielded a significantly higher  $H_t$  and  $H_s$  (i.e., less predictable gaze) than others. Jayawardena et al. (2020) performed entropy-based eye movement analysis on neurotypical and ADHD-diagnosed subjects during an audiovisual speech-in-noise task. They found that ADHD-diagnosed participants made unpredictable gaze transitions (i.e., high entropy) at different levels of task difficulty, whereas the neurotypical group of participants made gaze transitions from any AOI to the mouth region (i.e., low entropy) regardless of task difficulty. These findings suggest that  $H_t$  and  $H_s$  are potential indicators of curiosity, interest, picture familiarity, and task difficulty.

#### 4.6 Vergence Measures

Due to the association of ocular vergence with binocular vision, the binocular gaze data can be used to measure ocular vergence. The gaze vergence can be estimated (see **Figure 5**) using the distance between the individual gaze positions for each eye, the distance from the user to the screen, and the interocular distance (Daugherty et al., 2010; Wang et al., 2012). A common application of the ocular vergence is assessing stereoscopic perception (Essig et al., 2004; Daugherty et al., 2010; Wang et al., 2012).

These studies use ocular vergence to assess the user's depth perception under different stimuli conditions. Further, the ocular

vergence also has been the subject of estimating 3D gaze positions (Mlot et al., 2016) based on 2D positions provided by eye trackers.

The perceived depth ( $\Delta d$ ) can be computed using the distance between gaze positions ( $\Delta x$ ), the distance between the left and right eye ( $a$ ), and the distance between the user and the screen ( $D'$ ).

$$\Delta d = \frac{\Delta x D'}{\Delta x - a} \quad (4)$$

## 5 PUPIL MEASURES

Pupil measures (see **Table 1** for a list of pupil measures) capture fluctuations in the pupil's size and orientation to produce measurements that provide insights into one's internal state. The pupil diameter is the result of tonic and phasic pupillary responses of the eye (Sun et al., 1983; Wass et al., 2015; Peysakhovich et al., 2017).

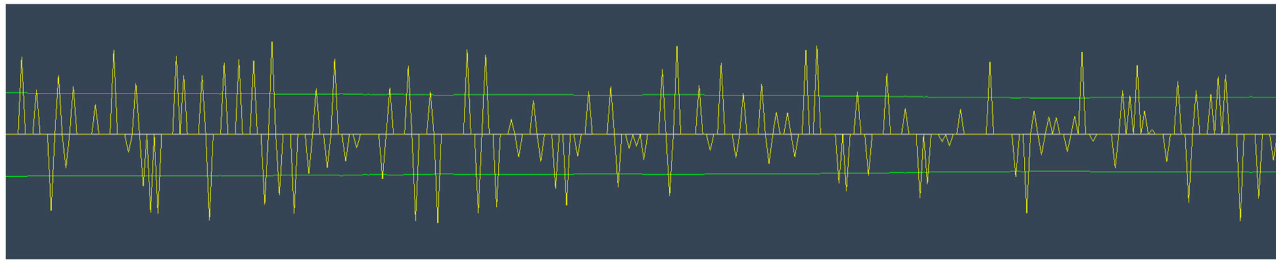
The tonic component refers to the pupil diameter changes caused by slow contractions, while the phasic refers to the quick or transient contractions. Most of the metrics use pupil dilation as the primary measure for the computations in order to determine the tonic and phasic components.

### 5.1 Pupil Diameter

The most primitive metric used in pupillometry is the average pupil diameter (Gray et al., 1993; Joshi et al., 2016). The measure captures both the tonic and the phasic components of the pupil dilation. Using the average dilation, some of the phasic and transient features can get smoothed out. An alternative measure to overcome the drawback of the average pupil dilation is to use the change in pupil dilation relative to a baseline. The baseline is assumed to correspond to the tonic component, while the relative change refers to the phasic component. The baseline can be either derived during the study or through a controlled environment.

### 5.2 Index of Cognitive Activity

Index of Cognitive Activity (ICA), introduced by Marshall (2002), is a measure of pupil diameter fluctuation as an instantaneous measure. Marshall (2000), Marshall (2007) describes the methodology followed in computing the ICA in an experiment setting. Furthermore, the publications also include parameter selections when performing experiments. The process starts by eliminating the pupil signal regions corresponding to blinks by either removing them or replacing them through interpolation as the preprocessing step. The signal is then passed through wavelet decomposition to capture the pupil signal's abrupt changes through decomposing to the desired level. Finally, the decomposed signal subjects to thresholding, converting the decomposed signal coefficients to a binary vector of the same size; the thresholding stage acts as a de-noising stage here.



**FIGURE 6 |** Visualization of the processed pupil diameter signal (yellow) and the threshold used for calculating the Low/High Index of Pupillary Activity (green) (Duchowski et al., 2020).

### 5.3 Index of Pupillary Activity

Index of Pupillary activity (IPA), introduced by Duchowski et al. (2018), is a metric inspired by ICA, with a similar underlying concept. Studies on ICA do not fully disclose the internals of ICA due to intellectual property reasons. IPA, however, discloses the internals of its process.

IPA computation starts by discarding pupil signals around blinks identified in the experiment. Duchowski et al. (2018) used a window of 200 ms in either direction during experiments. The procedure for computing the IPA measurement starts with a two-level Symlet-16 discrete wavelet decomposition of the pupil dilation signal by selecting a mother wavelet function  $\psi_{j,k}(t)$ . The resulting dyadic wavelet upon the wavelet analysis of the signal  $x(t)$ , generates a dyadic series representation. Then the process follows a multi-resolution signal analysis of the original signal  $x(t)$ . A level is arbitrarily selected from the multi-resolution decomposition to produce a smoother approximation of the signal  $x(t)$ . Finally, threshold the wavelet modulus maxima coefficients using a universal threshold defined by,

$$\lambda_{univ} = \hat{\sigma} \sqrt{2 \log n} \quad (5)$$

where  $\hat{\sigma}$  is the standard deviation of the noise. The number of remaining coefficients represents the IPA reading for the given pupil diameter signal.

### 5.4 Low/High Index of Pupillary Activity

Low/High Index of Pupillary Activity (LHIPA), introduced by Duchowski et al. (2020), is a variation of the IPA measure identified earlier. The computation of the metric remains the same as in IPA, except for counting remnants. Instead of counting threshold remnants as in IPA (Duchowski et al., 2018), LHIPA counts the modulus maxima of the low and high-frequency bands contained in the wavelet (see Figure 6).

Using the Discrete Wavelet Transform to analyze the pupil diameter signal at multiple levels of resolution, the wavelet coefficients are found by:

$$x_{\psi}^{(j-1)}(t) = \sum_k g_k x_{\psi}^j(2t + k) \quad (6)$$

where  $g_k$  is the one-dimensional high-pass wavelet filter. Level  $j$  is chosen arbitrarily to select either high or low frequency wavelet coefficients. For the high frequency component, they have chosen

$j = 1$  and for the low frequency component, they have chosen  $j = 1/2 \log_2(n)$ , the mid-level frequency octave where  $\log_2(n)$  is the number of octaves. Thus, the low frequency/high frequency ratio is:

$$x_{\psi}^{1/2 \log_2(n)}(t) / x_{\psi}^1(2^{1/2 \log_2(n)} t) \quad (7)$$

The pupillary response increases as the cognitive load increases. Since LHIPA is a ratio, an increase in the pupillary response reflects a decrease in the LHIPA reading. The authors demonstrate the metric's applicability using a series of three experiments where they assess the relationship between the task difficulty and the corresponding measures of IPA and LHIPA. During the experiments, the authors determined that the LHIPA was able to identify between difficult tasks and easy or baseline tasks throughout the study, while IPA could do so during only one experiment. The experiments also revealed that the LHIPA demonstrated cognitive load earlier than IPA, indicating a faster response due to the measure being a ratio and a built-in ratio arising from the ratio's computation.

## 6 DISCUSSION

Eye movement and pupil measures have been applied in disciplines including, but not limited to, neuroscience (Hessels and Hooze, 2019), psychology (Mele and Federici, 2012), and human computer interaction (Duchowski, 2002). In this section, we collectively summarize the literature of eye movement and pupillometry applications under different domains, discuss the limitations that we observe, and by doing so, establish a vision for implementing eye movements and pupil measures in these application environments.

### 6.1 Applications

#### 6.1.1 Neuroscience

In eye-tracking neuroscience, researchers have jointly analyzed neuronal activity and oculomotor activity to study the physiological organization of the vision system, and their effect on cognition and behavior. Studies of neuronal activity during phenomena such as attention (Blair et al., 2009; Kimble et al., 2010), scene perception (Duc et al., 2008), inattentive blindness (Simons and Chabris, 1999), visual engagement



(Catherine and James, 2001), and covert attention processing (i.e., visually fixating on one location while diverting attention to another) (Posner et al., 1980) had revealed important facts about cognition and behavior.

For instance, studies on covert attention processing show that attention cannot be inferred solely from whether an object was looked at (Hafed and Clark, 2002; Ebitz and Moore, 2019), and doing so would lead to false positives (Posner et al., 1980). Similarly, studies on scene perception show that information is processed in two forms: top-down (based on semantic importance) and bottom-up (based on visual significance such as color, brightness, etc.) (Duc et al., 2008). Moreover, studies on visual attention show that attention is divided among AOIs through a sustained cognitive state, from which relevant visual objects become available to influence behavior Duncan et al. (1994).

Studies have also revealed that pupillary activity is correlated with cognitive load (Hess and Polt, 1964; Hyönä et al., 1995), and also with neural gain (Eldar et al., 2013) and cortical activity (Reimer et al., 2016). While cognitive load can be inferred from pupillary activity, studies show that such inference becomes challenging in fast-paced cognitive tasks (Wierda et al., 2012), temporally overlapping cognitive tasks (Wierda et al., 2012), and in surroundings with varying ambient luminance (Zénon, 2017). Chong et al. (2020) showed that pupil diameter is regulated by sympathetic activation (arousal-induced pupil dilation) and parasympathetic inhibition (saccade-induced pupil dilation), both of which are affected by ambient luminance. Wierda et al. (2012) showed that a deconvolved pupil response signal is indicative of cognitive load, with a high temporal resolution. However, (Zénon, 2017), showed that this method does not account for low-frequency pupil fluctuations and inter-individual variability of pupil responses, and instead proposed using auto-regressive models (Cho et al., 2018) with exogenous inputs to analyze pupillary activity. Similarly, Watson and Yellott (2012) proposes a generalized formula to analyze pupillary activity, which accounts for ambient luminance, the size of the adapting field, the age of the observer, and whether both pupils are adapted.

Researchers of eye tracking neuroscience have utilized various measures to study neurodevelopmental disorders. Some of these studies are exclusively based on fixation measures. For instance, He et al. (2019) examined the *fixation count* and *fixation duration* in preschoolers with Autism Spectrum Disorder (ASD), and found that preschoolers with ASD had atypical gaze patterns in a facial emotion expression task compared to typical individuals. Moreover, they found that deficits in recognizing emotions from facial expressions in ASD correlated with the social interaction and development quotient. In contrast, some studies are based exclusively on saccadic measures. For instance, MacAskill et al. (2002) found that the *saccade amplitude* was modified during adaption of memory-guided saccades and this adaptive ability was impaired in individuals with Parkinson's disease. Similarly, Barbosa et al. (2019) identified that *saccadic direction* errors are associated with impulsive compulsive behaviors of individuals with Parkinson's Disease, as they had difficulty in suppressing automatic saccades to a given target.

Patel et al. (2012) showed that *saccade latency* is correlated with the severity of Huntington Disease (HD), and suggested its possibility of being a biomarker of disease severity in HD. In another study, Jensen et al. (2019) found that reduced *saccade velocity* is a key indicator of progressive supranuclear palsy and other disorders of mid-brain. Similarly, Biscaldi et al. (1998) found that individuals with Dyslexia had significantly higher regressive *saccade rate* in a sequential-target task. Similarly, Termsarasab et al. (2015) stated that abnormalities in *saccade gain* could aid diagnosis of hyperkinetic and hypokinetic movement disorders. Similarly, Fukushima et al. (2013) used a memory-based smooth pursuit task to examine working memory of *smooth pursuit direction* in individuals with PD.

Certain studies performed their analysis using visual search measures. Rutherford and Towns (2008), for instance, demonstrated *scan path* similarities and differences during emotion perception between typical individuals and individuals with ASD. Similarly, Bours et al. (2018) demonstrated that individuals with ASD had increased *time to first fixation* on the eyes of fearful faces during emotion recognition task. Duret et al. (1999) showed that *refixation* strategies in macular disorders was dependant on location of the target relative to the scotoma, spatial characteristics of the disease and the duration of the disorder. Guillon et al. (2015) showed that *gaze transition matrices* could potentially reveal new strategies of visual scanning followed by individuals with ASD. Moreover, Wainstein et al. (2017) showed that *pupil diameter* could be a biomarker in ADHD based on the results from a visuo-spatial working memory task.

### 6.1.2 Human Computer Interaction

In HCI research, eye tracking has been used to evaluate the usability of human-computer interfaces (both hardware and software). Here, the primary eye movement measures being analyzed are saccades, fixations, smooth pursuits, compensatory, vergence, micro-saccades, and nystagmus (Goldberg and Wichansky, 2003). For instance, Farbos et al. (2000) and Dobson (1977) had attempted to correlate eye tracking measures with software usability metrics such as time taken, completion rate, and other global metrics. Similarly, Du and MacDonald (2014) had analyzed how visual saliency is affected by icon size. In both scenarios, eye movements were recorded while users navigated human-computer interfaces, and were subsequently analyzed using fixation and scan-path measures.

Pupil diameter measures are widely used to assess the cognitive load of users when interacting with human-computer interfaces. For instance, Bailey and Iqbal (2008) used pupil diameter measures to demonstrate that cognitive load varies while completing a goal-directed task. Adamczyk and Bailey (2004) used pupil diameter measures to demonstrate that the disruption caused by user interface interruptions (e.g., notifications) is more pronounced during periods of high cognitive load. Iqbal et al. (2004) used pupil diameter measures to show that difficult tasks demand longer processing time, induces higher subjective ratings of cognitive load, and reliably evokes greater pupillary response at salient

subtasks. In addition to pupil diameter measures, studies such as Chen and Epps (2014) have also utilized blink rate measures to analyze how *cognitive load* (Sweller, 2011) and *perceptual load* (Macdonald and Lavie, 2011) varies across different tasks. They claim that pupil diameter indicates cognitive load well for tasks with low perceptual load, and that blink rate indicates perceptual load better than cognitive load.

### 6.1.3 Psychology

In psychology research, eye tracking has been used to understand eye movements and visual cognition during naturalistic interactions such as reading, driving, and speaking. Rayner (2012), for instance, have analyzed eye movements during English reading, and observed a mean saccade *duration* and *amplitude* of 200–250 ms and 7–9 letters, respectively. They have also observed different eye movement patterns when reading silently vs aloud, and correlations of text complexity to both *fixation duration* (+) and *saccade length* (–). Three paradigms were commonly used to analyze eye movements during reading tasks: *moving window*—selecting a few characters before/after the fixated word (McConkie and Rayner, 1975), *foveal mask*—masking a region around the fixated word (Bertera and Rayner, 2000), and *boundary*—creating pre-defined boundaries (i.e., AOIs) to classify fixations (Rayner, 1975). Over time, these paradigms have been extended to tasks such as scene perception. Recarte and Nunes (2000) and Stapel et al. (2020), for instance, have used the boundary method to analyze eye movements during driving tasks and thereby assess the effect of driver awareness on gaze behavior. In particular, Recarte and Nunes (2000) have observed that distracted drivers had higher *fixation durations*, higher *pupil dilations*, and lower *fixation counts* in the mirror/speedometer AOIs compared to control subjects.

In psycholinguistic research, pupil measures have been used to analyze the cognitive load of participants in tasks such as simultaneous interpretation (Russell, 2005), speech shadowing (Marslen-Wilson, 1985), and lexical translation. For instance, Seeber and Kerzel (2012) measured the pupil diameter during simultaneous interpretation tasks, and observed a larger average pupil diameter (indicating a higher cognitive load) when translating between verb-final and verb-initial languages, compared to translating between languages of the same type. Moreover, Hyönä et al. (1995) measured the pupil diameter during simultaneous interpretation, speech shadowing, and lexical translation tasks. They observed a larger average pupil diameter (indicating a higher cognitive load) during simultaneous interpretation than speech shadowing, and also momentary variations in pupil diameter (corresponding to spikes in cognitive load) during lexical translation.

Moreover, studies in marketing and behavioral finance, have used eye-tracking and pupillometric measures to understand the relationship between presented information and the decision-making process. For instance, Rubaltelli et al. (2016) used *pupil diameter* to understand investor decision-making, while Ceravolo et al. (2019), Hüsner and Wirth (2014) used *dwell time*. Further, studies in marketing have identified

relationships between the consumer decision process through fixations on different sections in the product description Ares et al. (2013), Menon et al. (2016). The utility of eye-tracking in marketing extends beyond product descriptions, to advertisements, brands, choice, and search patterns (Wedel and Pieters, 2008).

## 6.2 Recent Developments

Among the recent developments in eye tracking and pupil measures, the introduction of a series of pupillometry-based measures (ICA, IPA, and LHIPA) (Marshall, 2002; Duchowski et al., 2018; Duchowski et al., 2020) to assess the cognitive load is noteworthy. These measures, in general, compute the cognitive load by processing the pupil dilation as a signal, and thus require frequent and precise measurements of pupil dilation to function. We ascribe the success of these measures to the technological advancements in pupillometric devices, which led to higher sampling rates and more accurate measurement of pupil dilation than possible before. In the future, we anticipate the continued development of measures that exploit pupil dilation signals (El Haj and Moustafa, 2021; Maier and Grueschow, 2021).

Another noteworthy development is the emergence of commodity camera-based eye-tracking (Sewell and Komogortsev, 2010; Krafka et al., 2016; Semmelmann and Weigelt, 2018; Mahanama et al., 2020) systems to further democratize eye-tracking research and interaction. Since these systems require no dedicated/specialized hardware, one could explore eye-tracking at a significantly lower cost than otherwise possible. Yet the lack of specialized hardware, such as IR illumination or capture, restricts their capability to only eye tracking, and not pupillometry. The relatively low sampling rates of commodity cameras may negatively affect the quality of eye-tracking measures. Overall, eye tracking on commodity hardware provides a cost-effective means of incorporating other eye-tracking measures, despite its quality being heavily device-dependent. In the future, we anticipate cameras to have higher sampling frequencies (Wang et al., 2021) and resolutions, which, in turn, would bridge the gap between specialized eye trackers and commodity camera-based eye tracking systems.

## 6.3 Limitations

One of the major limitation we observed is that most studies use derivative measures of only a single oculomotor event instead of combinations of multiple events. For instance, studies that use fixational eye movements generally use only measures associated with fixations, despite the possible utility of saccadic information. This limitation is exacerbated with the confusion on the concepts of fixations and saccades (Hessels et al., 2018). Further, most studies rely on first-order statistical features (e.g., histogram-based features such as min, max, mean, median, sd, and skewness) or trend analysis (e.g., trajectory-based features such as sharpest decrease and sharpest increase between consecutive samples) on the features instead of

employing advanced measures. Our study identified some measures to be popular in specific domains, despite their potential applicability into other domains; for instance, pupillary measures are extensively studied in neuroscience research (Schwalm and Jubal, 2017; Schwiedrzik and Sudmann, 2020), but not quite so in psychology research. One plausible reason is the lack of literature that aggregates eye movement and pupillometric measures to assist researchers in identifying additional measures. Through this paper, we attempt to provide this missing knowledge to researchers. Another reason could be the computational limitations of eye tracking hardware and software. For instance, micro-saccadic measures often require high-frequency eye trackers ( $\geq 300$  Hz) (Krejtz et al., 2018) that are relatively expensive, thereby imposing hardware-level restrictions. Likewise, intellectual property restrictions (Marshall, 2000) and the lack of public implementations (Duchowski et al., 2018; Duchowski et al., 2020) (i.e., code libraries) of eye tracking solutions impose software-level restrictions. Both scenarios create an entry barrier for researchers into all applicable eye movement and pupillometry measures. Even though we suggest alternatives for patented measures in this paper, the lack of code libraries still remains unaddressed.

Another limitation is the relatively unexplored research avenues of eye-tracking and pupillometric measures in Extended Reality (XR) environments (Rappa et al., 2019; Renner and Pfeiffer, 2017; Clay et al., 2019; Mutasim et al., 2020; Heilmann and Witte, 2021), such as Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). These extended realities have a broad utility in simulating real-world scenarios, often eliminating the requirement of complex experimental environments (i.e., VR driver behavior (Bozkir et al., 2019)). Further, the significant control of the reality vested to the experiment designer enables them to simulate complex and rare real-world events. However, these extended reality scenarios require additional hardware and software to perform eye-tracking. This poses entry barriers for practitioners in the form of budgetary constraints (i.e., cost of additional hardware, and software), knowledge constraints (i.e., experience/knowledge on XR toolsets), and experimental design (i.e., the structure of the experiment, how to simulate and integrate modalities such as touch or haptics). We suspect these factors collectively contribute to the less exploratory studies in XR eye tracking research. Considering the possibility of this technology being more commonplace (e.g., Microsoft HoloLens 2<sup>1</sup> eye tracking optics, HTC VIVE Pro Eye<sup>2</sup>, VARJO VR eye tracking<sup>3</sup>), this could be an excellent opportunity for future studies.

## 7 CONCLUSION

In this paper, we have identified, discussed, and reviewed existing measures in eye-tracking and pupillometry. Further, we classified these measures based on the mechanism of vision; as eye movement, blink, and pupil-based measures. For each, we provided an overview of the measure, the underlying eye-mechanism exploited by it, and how it can be measured. Further, we identified a selected set of studies that use each type of measure and documented their findings.

This survey aims to help the researchers in two forms. First, due to the utility of eye-tracking and pupillometric measures in a broader range of domains, the implications of measures/results in other domains can easily be overlooked. Our study helps to overcome the issue by including applications and their indications along with each measure. Further, we believe the body of knowledge in the study would help researchers to choose appropriate measures for a future study. Researchers could adapt our taxonomy to classify eye-tracking and pupil measures based on the eye mechanism and vice versa. Moreover, the researchers can identify particular eye mechanisms or measures exploited for research through the presented classification. Finally, we expect this review to serve as a reference for researchers exploring eye tracking techniques using eye movements and pupil measures.

## AUTHOR CONTRIBUTIONS

BM: Planned the structure of the paper and content, carried out the review for the survey, organized the entire survey and contributed to writing the manuscript. YJ: Planned the structure of the paper and content, carried out the review for the survey, organized the entire survey and contributed to writing the manuscript. SR: Planned the structure of the paper and content, proofread, and contributed to writing the manuscript. GJ: Contributed to writing the Pupillometry measures. LC: Contributed to proofread, and supervision JS: Contributed to proofread, and supervision SJ: Idea formation, proofread and research supervision.

## FUNDING

This work is supported in part by the U.S. National Science Foundation grant CAREER IIS-2045523. Any opinions, findings and conclusion or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

## ACKNOWLEDGMENTS

We thank the reviewers whose comments/suggestions helped improve and clarify this manuscript.

<sup>1</sup><https://www.microsoft.com/en-us/hololens/hardware>

<sup>2</sup><https://www.vive.com/us/product/vive-pro-eye/overview/>

<sup>3</sup><https://varjo.com/use-center/get-to-know-your-headset/eye-tracking/>

## REFERENCES

- Abel, L. A., and Hertle, R. W. (1988). "Effects of Psychoactive Drugs on Ocular Motor Behavior," in *Neuropsychology of Eye Movements* (Hillsdale, NJ: Lawrence Erlbaum Associates), 81–114.
- Adamczyk, P. D., and Bailey, B. P. (2004). "If Not Now, when? the Effects of Interruption at Different Moments within Task Execution," in Proceedings of the SIGCHI conference on Human factors in computing systems, 271–278.
- Adler, F. H., and Fliegelman, M. (1934). Influence of Fixation on the Visual Acuity. *Arch. Ophthalmol.* 12, 475–483. doi:10.1001/archophth.1934.00830170013002
- Anderson, N. C., Bischof, W. F., Foulsham, T., and Kingstone, A. (2020). Turning the (Virtual) World Around: Patterns in Saccade Direction Vary with Picture Orientation and Shape in Virtual Reality. *J. Vis.* 20, 21. doi:10.1167/jov.20.8.21
- Andersson, R., Nyström, M., and Holmqvist, K. (2010). Sampling Frequency and Eye-Tracking Measures: How Speed Affects Durations, Latencies, and More. *J. Eye Move. Res.* 3. doi:10.16910/jemr.3.3.6
- Ansari, M. F., Kasprowski, P., and Obetkal, M. (2021). Gaze Tracking Using an Unmodified Web Camera and Convolutional Neural Network. *Appl. Sci.* 11, 9068. doi:10.3390/app11199068
- Anson, E. R., Bigelow, R. T., Carey, J. P., Xue, Q.-L., Studenski, S., Schubert, M. C., et al. (2016). Aging Increases Compensatory Saccade Amplitude in the Video Head Impulse Test. *Front. Neurol.* 7, 113. doi:10.3389/fneur.2016.00113
- Ares, G., Giménez, A., Bruzzone, F., Vidal, L., Antúnez, L., and Maiche, A. (2013). Consumer Visual Processing of Food Labels: Results from an Eye-Tracking Study. *J. Sens Stud.* 28, 138–153. doi:10.1111/joss.12031
- Bahill, A. T., Clark, M. R., and Stark, L. (1975). The Main Sequence, a Tool for Studying Human Eye Movements. *Math. biosciences* 24, 191–204. doi:10.1016/0025-5564(75)90075-9
- Bahill, A. T., and Laritz, T. (1984). Why Can't Batters Keep Their Eyes on the ball. *Am. Scientist* 72, 249–253.
- Bailey, B. P., and Iqbal, S. T. (2008). Understanding Changes in Mental Workload during Execution of Goal-Directed Tasks and its Application for Interruption Management. *ACM Trans. Comput.-Hum. Interact.* 14, 1–28. doi:10.1145/1314683.1314689
- Barbosa, P., Kaski, D., Castro, P., Lees, A. J., Warner, T. T., and Djamshidian, A. (2019). Saccadic Direction Errors Are Associated with Impulsive Compulsive Behaviours in Parkinson's Disease Patients. *Jpd* 9, 625–630. doi:10.3233/jpd-181460
- Barmack, N. H. (1970). Modification of Eye Movements by Instantaneous Changes in the Velocity of Visual Targets. *Vis. Res.* 10, 1431–1441. doi:10.1016/0042-6989(70)90093-3
- Barnes, G. R., and Asselman, P. T. (1991). The Mechanism of Prediction in Human Smooth Pursuit Eye Movements. *J. Physiol.* 439, 439–461. doi:10.1113/jphysiol.1991.sp018675
- Barnes, G. R. (2008). Cognitive Processes Involved in Smooth Pursuit Eye Movements. *Brain Cogn.* 68, 309–326. doi:10.1016/j.bandc.2008.08.020
- Bartels, M., and Marshall, S. P. (2012). "Measuring Cognitive Workload across Different Eye Tracking Hardware Platforms," in Proceedings of the symposium on eye tracking research and applications, 161–164. doi:10.1145/2168556.2168582
- Becker, W., and Fuchs, A. F. (1969). Further Properties of the Human Saccadic System: Eye Movements and Correction Saccades with and without Visual Fixation Points. *Vis. Res.* 9, 1247–1258. doi:10.1016/0042-6989(69)90112-6
- Bednarik, R., Myller, N., Sutinen, E., and Tukiainen, M. (2005). "Applying Eye-Movement Tracking to Program Visualization," in 2005 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'05) (IEEE), 302–304.
- Bertera, J. H., and Rayner, K. (2000). Eye Movements and the Span of the Effective Stimulus in Visual Search. *Perception & Psychophysics* 62, 576–585. doi:10.3758/bf03212109
- Birmingham, E., Bischof, W. F., and Kingstone, A. (2008). Social Attention and Real-World Scenes: The Roles of Action, Competition and Social Content. *Q. J. Exp. Psychol.* 61, 986–998. doi:10.1080/17470210701410375
- Biscaldi, M., Gezeck, S., and Stühr, V. (1998). Poor Saccadic Control Correlates with Dyslexia. *Neuropsychologia* 36, 1189–1202. doi:10.1016/s0028-3932(97)00170-x
- Blair, M. R., Watson, M. R., Walshe, R. C., and Maj, F. (2009). Extremely Selective Attention: Eye-Tracking Studies of the Dynamic Allocation of Attention to Stimulus Features in Categorization. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1196–1206. doi:10.1037/a0016272
- Blount, W. P. (1927). Studies of the Movements of the Eyelids of Animals: Blinking. *Exp. Physiol.* 18, 111–125. doi:10.1113/expphysiol.1927.sp000426
- Bojko, A. (2006). Using Eye Tracking to Compare Web page Designs: A Case Study. *J. Usability Stud.* 1, 112–120.
- Bours, C. C. A. H., Bakker-Huvenaars, M. J., Tramper, J., Bielczyk, N., Scheepers, F., Nijhof, K. S., et al. (2018). Emotional Face Recognition in Male Adolescents with Autism Spectrum Disorder or Disruptive Behavior Disorder: an Eye-Tracking Study. *Eur. Child. Adolesc. Psychiatry* 27, 1143–1157. doi:10.1007/s00787-018-1174-4
- Boxer, A. L., Garbutt, S., Seeley, W. W., Jafari, A., Heuer, H. W., Mirsky, J., et al. (2012). Saccade Abnormalities in Autopsy-Confirmed Frontotemporal Lobar Degeneration and Alzheimer Disease. *Arch. Neurol.* 69, 509–517. doi:10.1001/archneurol.2011.1021
- Bozkir, E., Geisler, D., and Kasneci, E. (2019). "Person Independent, Privacy Preserving, and Real Time Assessment of Cognitive Load Using Eye Tracking in a Virtual Reality Setup," in 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (IEEE), 1834–1837. doi:10.1109/vr.2019.8797758
- Braun, D. I., Pracejus, L., and Gegenfurtner, K. R. (2006). Motion Aftereffect Elicits Smooth Pursuit Eye Movements. *J. Vis.* 6, 1. doi:10.1167/6.7.1
- Březinová, V., and Kendell, R. (1977). Smooth Pursuit Eye Movements of Schizophrenics and normal People under Stress. *Br. J. Psychiatry* 130, 59–63.
- Brutten, G. J., and Janssen, P. (1979). An Eye-Marking Investigation of Anticipated and Observed Stuttering. *J. Speech Lang. Hear. Res.* 22, 20–28. doi:10.1044/jshr.2201.20
- Buonocore, A., Chen, C.-Y., Tian, X., Idrees, S., Münch, T. A., and Hafed, Z. M. (2017). Alteration of the Microsaccadic Velocity-Amplitude Main Sequence Relationship after Visual Transients: Implications for Models of Saccade Control. *J. Neurophysiol.* 117, 1894–1910. doi:10.1152/jn.00811.2016
- Buonocore, A., McIntosh, R. D., and Melcher, D. (2016). Beyond the point of No Return: Effects of Visual Distractors on Saccade Amplitude and Velocity. *J. Neurophysiol.* 115, 752–762. doi:10.1152/jn.00939.2015
- Burke, M. R., and Barnes, G. R. (2006). Quantitative Differences in Smooth Pursuit and Saccadic Eye Movements. *Exp. Brain Res.* 175, 596–608. doi:10.1007/s00221-006-0576-6
- Burr, D. C., Morrone, M. C., and Ross, J. (1994). Selective Suppression of the Magnocellular Visual Pathway during Saccadic Eye Movements. *Nature* 371, 511–513. doi:10.1038/371511a0
- Buswell, G. T. (1935). *How People Look at Pictures: A Study of the Psychology and Perception in Art*. Chicago: Univ. Chicago Press.
- Cardona, G., García, C., Serés, C., Vilaseca, M., and Gispets, J. (2011). Blink Rate, Blink Amplitude, and Tear Film Integrity during Dynamic Visual Display Terminal Tasks. *Curr. Eye Res.* 36, 190–197. doi:10.3109/02713683.2010.544442
- Carl, J. R., and Gellman, R. S. (1987). Human Smooth Pursuit: Stimulus-dependent Responses. *J. Neurophysiol.* 57, 1446–1463. doi:10.1152/jn.1987.57.5.1446
- Carpenter, R. H. (1988). *Movements of the Eyes*. 2nd Rev Pion Limited.
- Castello, E., Baroni, N., and Palleschini, E. (1998). Neurological and Auditory Brain Stem Response Findings in Human Immunodeficiency Virus-Positive Patients without Neurologic Manifestations. *Ann. Otol Rhinol Laryngol.* 107, 1054–1060. doi:10.1177/000348949810701210
- Catherine, L., and James, M. (2001). *Educating Children with Autism*. Washington, DC: National Academies Press.
- Ceder, A. (1977). Drivers' Eye Movements as Related to Attention in Simulated Traffic Flow Conditions. *Hum. Factors* 19, 571–581. doi:10.1177/001872087701900606
- Ceravolo, M. G., Farina, V., Fattobene, L., Leonelli, L., and Raggetti, G. (2019). Presentational Format and Financial Consumers' Behaviour: an Eye-Tracking Study. *Int. J. Bank Marketing* 37 (3), 821–837. doi:10.1108/ijbm-02-2018-0041
- Chen, S., and Epps, J. (2014). Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Human-Computer Interaction* 29, 390–413. doi:10.1080/07370024.2014.892428
- Cherang, Y. G., Baird, T., Chen, J. T., and Wang, C. A. (2020). Background Luminance Effects on Pupil Size Associated with Emotion and Saccade Preparation. *Sci. Rep.* 10, 1–11. doi:10.1038/s41598-020-72954-z
- Cho, S.-J., Brown-Schmidt, S., and Lee, W.-y. (2018). Autoregressive Generalized Linear Mixed Effect Models with Crossed Random Effects: An Application to



- Intensive Binary Time Series Eye-Tracking Data. *Psychometrika* 83, 751–771. doi:10.1007/s11336-018-9604-2
- Choi, J. E. S., Vaswani, P. A., and Shadmehr, R. (2014). Vigor of Movements and the Cost of Time in Decision Making. *J. Neurosci.* 34, 1212–1223. doi:10.1523/jneurosci.2798-13.2014
- Chu, C. A., Rosenfield, M., and Portello, J. K. (2014). Blink Patterns. *Optom. Vis. Sci.* 91, 297–302. doi:10.1097/oxp.0000000000000157
- Churchland, A. K., and Lisberger, S. G. (2002). Gain Control in Human Smooth-Pursuit Eye Movements. *J. Neurophysiol.* 87, 2936–2945. doi:10.1152/jn.2002.87.6.2936
- Clay, V., König, P., and König, S. (2019). Eye Tracking in Virtual Reality. *J. Eye Mov. Res.* 12. doi:10.16910/jemr.12.1.3
- Coeckelbergh, T. R. M., Brouwer, W. H., Cornelissen, F. W., Van Wolfelaar, P., and Kooijman, A. C. (2002). The Effect of Visual Field Defects on Driving Performance. *Arch. Ophthalmol.* 120, 1509–1516. doi:10.1001/archophth.120.11.1509
- Coëffé, C., and O’regan, J. K. (1987). Reducing the Influence of Non-target Stimuli on Saccade Accuracy: Predictability and Latency Effects. *Vis. Res.* 27, 227–240. doi:10.1016/0042-6989(87)90185-4
- Collewin, H., and Tamminga, E. P. (1984). Human Smooth and Saccadic Eye Movements during Voluntary Pursuit of Different Target Motions on Different Backgrounds. *J. Physiol.* 351, 217–250. doi:10.1113/jphysiol.1984.sp015242
- Connolly, J. D., Goodale, M. A., Goltz, H. C., and Munoz, D. P. (2005). Fmri Activation in the Human Frontal Eye Field Is Correlated with Saccadic Reaction Time. *J. Neurophysiol.* 94, 605–611. doi:10.1152/jn.00830.2004
- Cornsweet, T. N. (1956). Determination of the Stimuli for Involuntary Drifts and Saccadic Eye Movements\*. *J. Opt. Soc. Am.* 46, 987–993. doi:10.1364/josa.46.000987
- Costa, M., Simone, A., Vignali, V., Lantieri, C., and Palena, N. (2018). Fixation Distance and Fixation Duration to Vertical Road Signs. *Appl. Ergon.* 69, 48–57. doi:10.1016/j.apergo.2017.12.017
- Crevits, L., Simons, B., and Wildenbeest, J. (2003). Effect of Sleep Deprivation on Saccades and Eyelid Blinking. *Eur. Neurol.* 50, 176–180. doi:10.1159/000073060
- Daugherty, B. C., Duchowski, A. T., House, D. H., and Ramasamy, C. (2010). “Measuring Vergence over Stereoscopic Video with a Remote Eye Tracker,” in Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, 97–100. doi:10.1145/1743666.1743690
- Day, B. L., and Fitzpatrick, R. C. (2005). The Vestibular System. *Curr. Biol.* 15, R583–R586. doi:10.1016/j.cub.2005.07.053
- de Brouwer, S., Missal, M., and Lefèvre, P. (2001). Role of Retinal Slip in the Prediction of Target Motion during Smooth and Saccadic Pursuit. *J. Neurophysiol.* 86, 550–558. doi:10.1152/jn.2001.86.2.550
- De Brouwer, S., Yuksel, D., Blohm, G., Missal, M., and Lefèvre, P. (2002). What Triggers Catch-Up Saccades during Visual Tracking. *J. Neurophysiol.* 87, 1646–1650. doi:10.1152/jn.00432.2001
- De Bruin, J., Malan, K., and Eloff, J. (2013). “Saccade Deviation Indicators for Automated Eye Tracking Analysis,” in Proceedings of the 2013 conference on eye tracking south africa, 47–54. doi:10.1145/2509315.2509324
- de Hemptinne, C., Lefèvre, P., and Missal, M. (2006). Influence of Cognitive Expectation on the Initiation of Anticipatory and Visual Pursuit Eye Movements in the Rhesus Monkey. *J. Neurophysiol.* 95, 3770–3782. doi:10.1152/jn.00007.2006
- Dell’Osso, L. F., Van der Steen, J., Steinman, R. M., and Collewin, H. (1992). Foveation Dynamics in Congenital Nystagmus. II: Smooth Pursuit. *Doc Ophthalmol.* 79, 25–49. doi:10.1007/BF00160131
- Demberg, V., Kiagia, E., and Sayeed, A. (2013). The index of Cognitive Activity as a Measure of Linguistic Processing. *reading time* 500, 1500.
- Demberg, V. (2013). “Pupillometry: the index of Cognitive Activity in a Dual-Task Study,” in Proceedings of the Annual Meeting of the Cognitive Science Society, 2154–2159.35
- Dimigen, O., Valsecchi, M., Sommer, W., and Kliegl, R. (2009). Human Microsaccade-Related Visual Brain Responses. *J. Neurosci.* 29, 12321–12331. doi:10.1523/jneurosci.0911-09.2009
- Ditchburn, R. W., and Ginsborg, B. L. (1953). Involuntary Eye Movements during Fixation. *J. Physiol.* 119, 1–17. doi:10.1113/jphysiol.1953.sp004824
- Dobson, M. W. (1977). Eye Movement Parameters and Map reading. *The Am. Cartographer* 4, 39–58. doi:10.1559/15230407784080022
- Dodge, R. (1903). Five Types of Eye Movement in the Horizontal meridian Plane of the Field of Regard. *Am. J. physiology-legacy content* 8, 307–329. doi:10.1152/ajplegacy.1903.8.4.307
- Dodge, R. (1900). Visual Perception during Eye Movement. *Psychol. Rev.* 7, 454–465. doi:10.1037/h0067215
- Donovan, T., and Litchfield, D. (2013). Looking for Cancer: Expertise Related Differences in Searching and Decision Making. *Appl. Cognit. Psychol.* 27, 43–49. doi:10.1002/acp.2869
- Doughty, M. J. (2001). Consideration of Three Types of Spontaneous Eyeblick Activity in normal Humans: during reading and Video Display Terminal Use, in Primary Gaze, and while in Conversation. *Optom. Vis. Sci.* 78, 712–725. doi:10.1097/00006324-200110000-00011
- Doughty, M. J. (2019). Effect of Distance Vision and Refractive Error on the Spontaneous Eye Blink Activity in Human Subjects in Primary Eye Gaze. *J. Optom.* 12, 111–119. doi:10.1016/j.optom.2018.03.004
- Doughty, M. J. (2002). Further Assessment of Gender- and Blink Pattern-Related Differences in the Spontaneous Eyeblick Activity in Primary Gaze in Young Adult Humans. *Optom. Vis. Sci.* 79, 439–447. doi:10.1097/00006324-200207000-00013
- Doughty, M. J., and Naase, T. (2006). Further Analysis of the Human Spontaneous Eye Blink Rate by a Cluster Analysis-Based Approach to Categorize Individuals with ‘Normal’ versus ‘Frequent’ Eye Blink Activity. *Eye & contact lens* 32, 294–299. doi:10.1097/01.icl.0000224359.32709.4d
- Du, P., and MacDonald, E. F. (2014). Eye-tracking Data Predict Importance of Product Features and Saliency of Size Change. *J. Mech. Des.* 136. doi:10.1115/1.4027387
- Duchowski, A. T. (2002). A Breadth-First Survey of Eye-Tracking Applications. *Behav. Res. Methods Instr. Comput.* 34, 455–470. doi:10.3758/BF03195475
- Duchowski, A. T. (2017). *Eye Tracking Methodology: Theory and Practice*. Springer.
- Duchowski, A. T., Krejtz, K., Gehrer, N. A., Bafna, T., and Bækgaard, P. (2020). “The Low/high index of Pupillary Activity,” in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12. doi:10.1145/3313831.3376394
- Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., et al. (2018). “The index of Pupillary Activity: Measuring Cognitive Load Vis-À-Vis Task Difficulty with Pupil Oscillation,” in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–13.
- Duncan, J., Ward, R., and Shapiro, K. (1994). Direct Measurement of Attentional Dwell Time in Human Vision. *Nature* 369, 313–315. doi:10.1038/369313a0
- Duret, F., Buquet, C., Charlier, J., Mermoud, C., Viviani, P., and Safran, A. B. (1999). Refixation Strategies in Four Patients with Macular Disorders. *Neuro-ophthalmology* 22, 209–220. doi:10.1076/noph.22.4.209.3718
- Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., and Tanenhaus, M. K. (1995). Eye Movements as a Window into Real-Time Spoken Language Comprehension in Natural Contexts. *J. Psycholinguist Res.* 24, 409–436. doi:10.1007/bf02143160
- Ebitz, R. B., and Moore, T. (2019). Both a Gauge and a Filter: Cognitive Modulations of Pupil Size. *Front. Neurol.* 9, 1190. doi:10.3389/fneur.2018.01190
- El Haj, M., and Moustafa, A. A. (2021). Pupil Dilation as an Indicator of Future Thinking. *Neurol. Sci.* 42, 647–653. doi:10.1007/s10072-020-04533-z
- Eldar, E., Cohen, J. D., and Niv, Y. (2013). The Effects of Neural Gain on Attention and Learning. *Nat. Neurosci.* 16, 1146–1153. doi:10.1038/nn.3428
- Ellis, S., Candrea, R., Misner, J., Craig, C. S., Lankford, C. P., and Hutchinson, T. E. (1998). “Windows to the Soul? what Eye Movements Tell Us about Software Usability,” in Proceedings of the usability professionals’ association conference, 151–178.
- Essig, K., Pomplun, M., and Ritter, H. (2004). “Application of a Novel Neural Approach to 3d Gaze Tracking,” in *Vergence Eye-Movements in Autostereograms*, 26. eScholarship.
- Ettinger, U., Kumari, V., Chitnis, X. A., Corr, P. J., Sumich, A. L., Rabe-Hesketh, S., et al. (2002). Relationship between Brain Structure and Saccadic Eye Movements in Healthy Humans. *Neurosci. Lett.* 328, 225–228. doi:10.1016/s0304-3940(02)00517-7
- Farbos, B., Mollard, R., Cabon, P., and David, H. (2000). “Measurement of Fatigue and Adaptation in Large-Scale Real-Time Atc Simulation,” in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA (Los Angeles, CA: SAGE Publications Sage CA), 3–204. doi:10.1177/154193120004401905

- Fehrer, B. C. O. F. (2021). Optimizing the Usage of Pupillary Based Indicators for Cognitive Workload. *J. Eye Mov. Res.* 14. doi:10.16910/jemr.14.2.4
- Fehrer, B. C. (2020). "One Threshold to Rule Them All? Modification of the index of Pupillary Activity to Optimize the Indication of Cognitive Load," in ACM Symposium on Eye Tracking Research and Applications, 1–5.
- Findlay, J. M., Walker, R., and Kentridge, R. W. (1995). *Eye Movement Research: Mechanisms, Processes and Applications*. Elsevier.
- Fischer, B., and Breitmeyer, B. (1987). Mechanisms of Visual Attention Revealed by Saccadic Eye Movements. *Neuropsychologia* 25, 73–83. doi:10.1016/0028-3932(87)90044-3
- Fischer, B., and Weber, H. (1993). Express Saccades and Visual Attention. *Behav. Brain Sci.* 16, 553–567. doi:10.1017/s0140525x00031575
- Fisher, D. F., Monty, R. A., and Senders, J. W. (2017). *Eye Movements: Cognition and Visual Perception*, Vol. 8. Oxfordshire, England, UK: Routledge.
- Foulsham, T., Kingstone, A., and Underwood, G. (2008). Turning the World Around: Patterns in Saccade Direction Vary with Picture Orientation. *Vis. Res.* 48, 1777–1790. doi:10.1016/j.visres.2008.05.018
- Fukushima, K., Fukushima, J., Warabi, T., and Barnes, G. R. (2013). Cognitive Processes Involved in Smooth Pursuit Eye Movements: Behavioral Evidence, Neural Substrate and Clinical Correlation. *Front. Syst. Neurosci.* 7, 4. doi:10.3389/fnsys.2013.00004
- Galley, N., Schleicher, R., and Galley, L. (2004). Blink Parameters as Indicators of Driver's Sleepiness—Possibilities and Limitations. *Vis. vehicles* 10, 189–196.
- Galley, N. (1993). The Evaluation of the Electrooculogram as a Psychophysiological Measuring Instrument in the Driver Study of Driver Behaviour. *Ergonomics* 36, 1063–1070. doi:10.1080/00140139308967978
- Gbadamosi, J., and Zangemeister, W. H. (2001). Visual Imagery in Hemianopic Patients. *J. Cogn. Neurosci.* 13, 855–866. doi:10.1162/089892901753165782
- Goldberg, J. H., and Helfman, J. I. (2010). "Scanpath Clustering and Aggregation," in Proceedings of the 2010 symposium on eye-tracking research & applications, 227–234. doi:10.1145/1743666.1743721
- Goldberg, J. H., and Kotval, X. P. (1999). Computer Interface Evaluation Using Eye Movements: Methods and Constructs. *Int. J. Ind. Ergon.* 24, 631–645. doi:10.1016/s0169-8141(98)00068-7
- Goldberg, J. H., and Wichansky, A. M. (2003). "Eye Tracking in Usability Evaluation," in *The Mind's Eye*. Editors J. Hyönä, R. Radach, and H. Deubel (Amsterdam: North-Holland), 493–516. doi:10.1016/B978-044451020-4/50027-X
- Goldberg, J. M., and Fernández, C. (1984). The Vestibular System. *Handbook Physiol.* 3, 977–1022. doi:10.1002/cphy.cp010321
- Goldberg, M. E., Eggers, H., and Gouras, P. (1991). The Oculomotor System. *Principles Neural Sci.*, 660–676.
- Gray, L. S., Winn, B., and Gilmartin, B. (1993). Accommodative Microfluctuations and Pupil Diameter. *Vis. Res.* 33, 2083–2090. doi:10.1016/0042-6989(93)90007-j
- Griffiths, A., Marshall, R., and Richens, A. (1984). Saccadic Eye Movement Analysis as a Measure of Drug Effects on Human Psychomotor Performance. *Br. J. Clin. Pharmacol.* 18, 73S–82S. doi:10.1111/j.1365-2125.1984.tb02584.x
- Grindinger, T. J., Murali, V. N., Tetreault, S., Duchowski, A. T., Birchfield, S. T., and Orero, P. (2010). "Algorithm for Discriminating Aggregate Gaze Points: Comparison with Salient Regions-Of-Interest," in Asian Conference on Computer Vision (Springer), 390–399.
- Guillon, Q., Afzali, M. H., Rogé, B., Baduel, S., Kruck, J., and Hadjikhani, N. (2015). The Importance of Networking in Autism Gaze Analysis. *PLoS one* 10, e0141191. doi:10.1371/journal.pone.0141191
- Guo, F., Ding, Y., Liu, W., Liu, C., and Zhang, X. (2016). Can Eye-Tracking Data Be Measured to Assess Product Design?: Visual Attention Mechanism Should Be Considered. *Int. J. Ind. Ergon.* 53, 229–235. doi:10.1016/j.ergon.2015.12.001
- Hafed, Z. M., Chen, C.-Y., and Tian, X. (2015). Vision, Perception, and Attention through the Lens of Microsaccades: Mechanisms and Implications. *Front. Syst. Neurosci.* 9, 167. doi:10.3389/fnsys.2015.00167
- Hafed, Z. M., and Clark, J. J. (2002). Microsaccades as an Overt Measure of covert Attention Shifts. *Vis. Res.* 42, 2533–2545. doi:10.1016/s0042-6989(02)00263-8
- Hauland, G., and Duijm, N. (2002). "Eye Movement Based Measures of Team Situation Awareness (Tsa)," in *Japan-halden MMS Workshop* (Kyoto, Japan: Kyoto University), 82–85.
- He, Y., Su, Q., Wang, L., He, W., Tan, C., Zhang, H., et al. (2019). The Characteristics of Intelligence Profile and Eye Gaze in Facial Emotion Recognition in Mild and Moderate Preschoolers with Autism Spectrum Disorder. *Front. Psychiatry* 10, 402. doi:10.3389/fpsyt.2019.00402
- Heilmann, F., and Witte, K. (2021). Perception and Action under Different Stimulus Presentations: A Review of Eye-Tracking Studies with an Extended View on Possibilities of Virtual Reality. *Appl. Sci.* 11, 5546. doi:10.3390/app11125546
- Hejtmančík, J. F., Cabrera, P., Chen, Y., M'Hamdi, O., and Nickerson, J. M. (2017). "Vision," in *Conn's Translational Neuroscience*. Editor P. M. Conn (San Diego: Academic Press), 399–438. doi:10.1016/B978-0-12-802381-5.00031-2
- Hemingshous, J., and Duchowski, A. T. (2006). "Icomp: a Tool for Scanpath Visualization and Comparison," in Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization, 152.
- Henderson, J. M., Choi, W., Luke, S. G., and Schmidt, J. (2018). Neural Correlates of Individual Differences in Fixation Duration during Natural reading. *Q. J. Exp. Psychol.* 71, 314–323. doi:10.1080/17470218.2017.1329322
- Hess, E. H., and Polt, J. M. (1964). Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science* 143, 1190–1192. doi:10.1126/science.143.3611.1190
- Hessels, R. S., and Hooge, I. T. C. (2019). Eye Tracking in Developmental Cognitive Neuroscience - the Good, the Bad and the Ugly. *Dev. Cogn. Neurosci.* 40, 100710. doi:10.1016/j.dcn.2019.100710
- Hessels, R. S., Kemner, C., van den Boomen, C., and Hooge, I. T. C. (2016). The Area-Of-Interest Problem in Eyetracking Research: A Noise-Robust Solution for Face and Sparse Stimuli. *Behav. Res.* 48, 1694–1712. doi:10.3758/s13428-015-0676-y
- Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., and Hooge, I. T. C. (2018). Is the Eye-Movement Field Confused about Fixations and Saccades? A Survey Among 124 Researchers. *R. Soc. Open Sci.*, 180502. doi:10.1098/rsos.180502
- Hoang Duc, A., Bays, P., and Husain, M. (2008). Eye Movements as a Probe of Attention. *Prog. Brain Res.* 171, 403–411. doi:10.1016/s0079-6123(08)00659-6
- Holland, C., and Komogortsev, O. V. (2011). "Biometric Identification via Eye Movement Scanpaths in reading," in 2011 International joint conference on biometrics (IJCB), 1–8. doi:10.1109/ijcb.2011.6117536IEEE
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and Van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford, UK: OUP Oxford.
- Horley, K., Williams, L. M., Gonsalvez, C., and Gordon, E. (2003). Social Phobics Do Not See Eye to Eye. *J. anxiety Disord.* 17, 33–44. doi:10.1016/s0887-6185(02)00180-9
- Hubel, D. H. (1995). *Eye, Brain, and Vision*. Scientific American Library/Scientific American Books.
- Hubel, D. H., and Wiesel, T. N. (1979). Brain Mechanisms of Vision. *Sci. Am.* 241, 150–162. doi:10.1038/scientificamerican0979-150
- Hüsler, A., and Wirth, W. (2014). Do investors Show an Attentional Bias toward Past Performance? an Eye-Tracking experiment on Visual Attention to Mutual Fund Disclosures in Simplified Fund Prospectuses. *J. Financ. Serv. Mark* 19, 169–185. doi:10.1057/fsm.2014.20
- Hyönä, J., Tömmola, J., and Alaja, A.-M. (1995). Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. *The Q. J. Exp. Psychol. Section A* 48, 598–612. doi:10.1080/14640749508401407
- Iqbal, S. T., Zheng, X. S., and Bailey, B. P. (2004). "Task-evoked Pupillary Response to Mental Workload in Human-Computer Interaction," in *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, 1477–1480. doi:10.1145/985921.986094
- Jacob, R. J. K., and Karn, K. S. (2003). "Eye Tracking in Human-Computer Interaction and Usability Research," in *The Mind's Eye*. Editors J. Hyönä, R. Radach, and H. Deubel (Amsterdam: North-Holland), 573–605. doi:10.1016/B978-044451020-4/50031-1
- Jainta, S., Vernet, M., Yang, Q., and Kapoula, Z. (2011). The Pupil Reflects Motor Preparation for Saccades - Even before the Eye Starts to Move. *Front. Hum. Neurosci.* 5, 97. doi:10.3389/fnhum.2011.00097
- Jarodzka, H., Holmqvist, K., and Nyström, M. (2010). "A Vector-Based, Multidimensional Scanpath Similarity Measure," in Proceedings of the 2010 symposium on eye-tracking research & applications, 211–218. doi:10.1145/1743666.1743718

- Jayawardena, G., and Jayarathna, S. (2021). "Automated Filtering of Eye Movements Using Dynamic Aoi in Multiple Granularity Levels," in *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 12, 49–64. doi:10.4018/ijmDEM.2021010104
- Jayawardena, G., Michalek, A., Duchowski, A., and Jayarathna, S. (2020). "Pilot Study of Audiovisual Speech-In-Noise (Sin) Performance of Young Adults with Adhd," in *ACM Symposium on Eye Tracking Research and Applications*, 1–5. doi:10.1145/3379156.3391373
- Jensen, K., Beylergil, S. B., and Shaikh, A. G. (2019). Slow Saccades in Cerebellar Disease. *cerebellum ataxias* 6, 1–9. doi:10.1186/s40673-018-0095-9
- Jiang, M.-Q. (1996). *The Role of Attention Mechanisms in Smooth Pursuit Performance in normal and Schizophrenic Subjects*. Iowa City, IA: The University of Iowa.
- Jongkees, B. J., and Colzato, L. S. (2016). Spontaneous Eye Blink Rate as Predictor of Dopamine-Related Cognitive Function-A Review. *Neurosci. Biobehavioral Rev.* 71, 58–82. doi:10.1016/j.neubiorev.2016.08.020
- Joshi, S., Li, Y., Kalwani, R. M., and Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron* 89, 221–234. doi:10.1016/j.neuron.2015.11.028
- Kao, G. W., and Morrow, M. J. (1994). The Relationship of Anticipatory Smooth Eye Movement to Smooth Pursuit Initiation. *Vis. Res.* 34, 3027–3036. doi:10.1016/0042-6989(94)90276-3
- Karsh, R., and Breitenbach, F. (1983). Looking at Looking: The Amorphous Fixation Measure. *Eye movements Psychol. functions: Int. views*, 53–64. doi:10.4324/9781003165538-6
- Killian, N. J., Potter, S. M., and Buffalo, E. A. (2015). Saccade Direction Encoding in the Primate Entorhinal Cortex during Visual Exploration. *Proc. Natl. Acad. Sci. USA* 112, 15743–15748. doi:10.1073/pnas.1417059112
- Kimble, M. O., Fleming, K., Bandy, C., Kim, J., and Zambetti, A. (2010). Eye Tracking and Visual Attention to Threatening Stimuli in Veterans of the Iraq War. *J. anxiety Disord.* 24, 293–299. doi:10.1016/j.janxdis.2009.12.006
- Knox, P. C., and Wolohan, F. D. A. (2014). Cultural Diversity and Saccade Similarities: Culture Does Not Explain Saccade Latency Differences between Chinese and Caucasian Participants. *PloS one* 9, e94424. doi:10.1371/journal.pone.0094424
- Komogortsev, O., Holland, C., Jayarathna, S., and Karpov, A. (2013). 2D Linear Oculomotor Plant Mathematical Model. *ACM Trans. Appl. Percept.* 10, 1–18. doi:10.1145/2536764.2536774
- Komogortsev, O. V., Jayarathna, S., Koh, D. H., and Gowda, S. M. (2010). "Qualitative and Quantitative Scoring and Evaluation of the Eye Movement Classification Algorithms," in *Proceedings of the 2010 Symposium on eye-tracking research & applications*, 65–68. doi:10.1145/1743666.1743682
- Korbach, A., Brünken, R., and Park, B. (2018). Differentiating Different Types of Cognitive Load: A Comparison of Different Measures. *Educ. Psychol. Rev.* 30, 503–529. doi:10.1007/s10648-017-9404-8
- Korbach, A., Brünken, R., and Park, B. (2017). Measurement of Cognitive Load in Multimedia Learning: a Comparison of Different Objective Measures. *Instr. Sci.* 45, 515–536. doi:10.1007/s11251-017-9413-5
- Koster, W. (1895). Étude sur les cônes et les bâtonnets dans la region de la fovea centralis de la rétine chez l'homme. *Arch. D'opt.* 15, 428.
- Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., et al. (2016). "Eye Tracking for Everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2176–2184. doi:10.1109/cvpr.2016.239
- Krauzlis, R. J. (2004). Recasting the Smooth Pursuit Eye Movement System. *J. Neurophysiol.* 91, 591–603. doi:10.1152/jn.00801.2003
- Krejtz, K., Żurawska, J., Duchowski, A. T., and Wichary, S. (2020). Pupillary and Microsaccadic Responses to Cognitive Effort and Emotional Arousal during Complex Decision Making. *J. Eye Mov Res.* 13. doi:10.16910/jemr.13.5.2
- Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., González Perilli, F., Pires, A., et al. (2015). Gaze Transition Entropy. *ACM Trans. Appl. Percept.* 13, 1–20. doi:10.1145/2834121
- Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., and Krejtz, I. (2018). Eye Tracking Cognitive Load Using Pupil Diameter and Microsaccades with Fixed Gaze. *PloS one* 13, e0203629. doi:10.1371/journal.pone.0203629
- Krejtz, K., Szmidt, T., Duchowski, A. T., and Krejtz, I. (2014). "Entropy-based Statistical Analysis of Eye Movement Transitions," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 159–166. doi:10.1145/2578153.2578176
- Krupinski, E. A. (1996). Visual Scanning Patterns of Radiologists Searching Mammograms. *Acad. Radiol.* 3, 137–144. doi:10.1016/s1076-6332(05)80381-2
- Ladda, J., Eggert, T., Glasauer, S., and Straube, A. (2007). Velocity Scaling of Cue-Induced Smooth Pursuit Acceleration Obeys Constraints of Natural Motion. *Exp. Brain Res.* 182, 343–356. doi:10.1007/s00221-007-0988-y
- Lai, H. Y., Saavedra-Pena, G., Sodini, C. G., Sze, V., and Heldt, T. (2019). Measuring Saccade Latency Using Smartphone Cameras. *IEEE J. Biomed. Health Inform.* 24, 885–897. doi:10.1109/JBHI.2019.2913846
- Land, M. F., and Furneaux, S. (1997). The Knowledge Base of the Oculomotor System. *Phil. Trans. R. Soc. Lond. B* 352, 1231–1239. doi:10.1098/rstb.1997.0105
- Le Meur, O., Coutrot, A., Liu, Z., Rämä, P., Le Roch, A., and Helo, A. (2017). Visual Attention Saccadic Models Learn to Emulate Gaze Patterns from Childhood to Adulthood. *IEEE Trans. Image Process.* 26, 4777–4789. doi:10.1109/tip.2017.2722238
- Lee, K. I., Jeon, J. H., and Song, B. C. (2020). "Deep Learning-Based Pupil center Detection for Fast and Accurate Eye Tracking System," in *European Conference on Computer Vision (Springer)*, 36–52. doi:10.1007/978-3-030-58529-7\_3
- Lehtinen, I., Lang, A. H., Jäntti, V., and Keskinen, E. (1979). Acute Effects of Alcohol on Saccadic Eye Movements. *Psychopharmacology* 63, 17–23. doi:10.1007/bf00426915
- Lisberger, S. G., Morris, E., and Tychsen, L. (1987). Visual Motion Processing and Sensory-Motor Integration for Smooth Pursuit Eye Movements. *Annu. Rev. Neurosci.* 10 (1), 97–129. doi:10.1146/annurev.ne.10.030187.000525
- Lisi, M., Solomon, J. A., and Morgan, M. J. (2019). Gain Control of Saccadic Eye Movements Is Probabilistic. *Proc. Natl. Acad. Sci. USA* 116, 16137–16142. doi:10.1073/pnas.1901963116
- MacAskill, M. R., Anderson, T. J., and Jones, R. D. (2002). Adaptive Modification of Saccade Amplitude in Parkinson's Disease. *Brain* 125, 1570–1582. doi:10.1093/brain/awf168
- Macdonald, J. S. P., and Lavie, N. (2011). Visual Perceptual Load Induces Inattentional Deafness. *Atten Percept Psychophys* 73, 1780–1789. doi:10.3758/s13414-011-0144-4
- Maffei, A., and Angrilli, A. (2019). Spontaneous Blink Rate as an index of Attention and Emotion during Film Clips Viewing. *Physiol. Behav.* 204, 256–263. doi:10.1016/j.physbeh.2019.02.037
- Mahanama, B., Jayawardana, Y., and Jayarathna, S. (2020). "Gaze-net: Appearance-Based Gaze Estimation Using Capsule Networks," in *Proceedings of the 11th Augmented Human International Conference*, 1–4.
- Maier, J. X., and Groh, J. M. (2009). Multisensory Guidance of Orienting Behavior. *Hearing Res.* 258, 106–112. doi:10.1016/j.heares.2009.05.008
- Maier, S. U., and Grueschow, M. (2021). Pupil Dilation Predicts Individual Self-Regulation success across Domains. *Sci. Rep.* 11, 1–18. doi:10.1038/s41598-021-93121-y
- Mania, K., McNamara, A., and Polychronakis, A. (2021). "Gaze-aware Displays and Interaction," in *ACM SIGGRAPH 2021 Courses*, 1–67. doi:10.1145/3450508.3464606
- Marshall, S. P. (2007). Identifying Cognitive State from Eye Metrics. *Aviat Space Environ. Med.* 78, B165–B175.
- Marshall, S. P. (2000). Method and Apparatus for Eye Tracking and Monitoring Pupil Dilation to Evaluate Cognitive Activity. *US Patent* 6, 051–090. [Dataset].
- Marshall, S. P. (2002). "The index of Cognitive Activity: Measuring Cognitive Workload," in *Proceedings of the IEEE 7th conference on Human Factors and Power Plants (IEEE)*, 7.
- Marslen-Wilson, W. D. (1985). Speech Shadowing and Speech Comprehension. *Speech Commun.* 4, 55–73. doi:10.1016/0167-6393(85)90036-6
- Martinez-Conde, S., Macknik, S. L., and Hubel, D. H. (2004). The Role of Fixational Eye Movements in Visual Perception. *Nat. Rev. Neurosci.* 5, 229–240. doi:10.1038/nrn1348
- May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., and Brannan, J. R. (1990). Eye Movement Indices of Mental Workload. *Acta psychologica* 75, 75–89. doi:10.1016/0001-6918(90)90067-p
- McConkie, G. W., and Rayner, K. (1975). The Span of the Effective Stimulus during a Fixation in reading. *Perception & Psychophysics* 17, 578–586. doi:10.3758/bf03203972
- McGregor, D. K., and Stern, J. A. (1996). Time on Task and Blink Effects on Saccade Duration. *Ergonomics* 39, 649–660. doi:10.1080/00140139608964487



- McKee, S. P., Levi, D. M., Schor, C. M., and Movshon, J. A. (2016). Saccadic Latency in Amblyopia. *J. Vis.* 16, 3. doi:10.1167/16.5.3
- McSorley, E., McCloy, R., and Lyne, C. (2012). The Spatial Impact of Visual Distractors on Saccade Latency. *Vis. Res.* 60, 61–72. doi:10.1016/j.visres.2012.03.007
- Megaw, E. D. (1979). Factors Affecting Visual Inspection Accuracy. *Appl. Ergon.* 10, 27–32. doi:10.1016/0003-6870(79)90006-1
- Megaw, E. D., and Richardson, J. (1979). Eye Movements and Industrial Inspection. *Appl. Ergon.* 10, 145–154. doi:10.1016/0003-6870(79)90138-8
- Meghanathan, R. N., Nikolaev, A. R., and van Leeuwen, C. (2019). Refixation Patterns Reveal Memory-Encoding Strategies in Free Viewing. *Atten Percept Psychophys* 81, 2499–2516. doi:10.3758/s13414-019-01735-2
- Mele, M. L., and Federici, S. (2012). Gaze and Eye-Tracking Solutions for Psychological Research. *Cogn. Process.* 13, 261–265. doi:10.1007/s10339-012-0499-z
- Mello-Thoms, C., Hardesty, L., Sumkin, J., Ganott, M., Hakim, C., Britton, C., et al. (2005). Effects of Lesion Conspicuity on Visual Search in Mammogram reading. *Acad. Radiol.* 12, 830–840. doi:10.1016/j.acra.2005.03.068
- Menon, R. G. V., Sigurdsson, V., Larsen, N. M., Fagerström, A., and Foxall, G. R. (2016). Consumer Attention to price in Social Commerce: Eye Tracking Patterns in Retail Clothing. *J. Business Res.* 69, 5008–5013. doi:10.1016/j.jbusres.2016.04.072
- Meyer, C. H., Lasker, A. G., and Robinson, D. A. (1985). The Upper Limit of Human Smooth Pursuit Velocity. *Vis. Res.* 25, 561–563. doi:10.1016/0042-6989(85)90160-9
- Michell, A. W., Xu, Z., Fritz, D., Lewis, S. J. G., Foltynie, T., Williams-Gray, C. H., et al. (2006). Saccadic Latency Distributions in Parkinson's Disease and the Effects of L-Dopa. *Exp. Brain Res.* 174, 7–18. doi:10.1007/s00221-006-0412-z
- Mlot, E. G., Bahmani, H., Wahl, S., and Kasneci, E. (2016). “3d Gaze Estimation Using Eye Vergence,” in Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (Rome, Italy: HEALTHINF), 125–131. doi:10.5220/0005821201250131
- Mohanty, A., and Sussman, T. J. (2013). Top-down Modulation of Attention by Emotion. *Front. Hum. Neurosci.* 7, 102. doi:10.3389/fnhum.2013.00102
- Morris, T. L., and Miller, J. C. (1996). Electrooculographic and Performance Indices of Fatigue during Simulated Flight. *Biol. Psychol.* 42, 343–360. doi:10.1016/0301-0511(95)05166-x
- Mostofi, N., Zhao, Z., Intoy, J., Boi, M., Victor, J. D., and Rucci, M. (2020). Spatiotemporal Content of Saccade Transients. *Curr. Biol.* 30, 3999–4008. doi:10.1016/j.cub.2020.07.085
- Motoki, K., Saito, T., and Onuma, T. (2021). Eye-tracking Research on Sensory and Consumer Science: A Review, Pitfalls and Future Directions. *Food Res. Int.* 145, 110389. doi:10.1016/j.foodres.2021.110389
- Mulder, K., Klugkist, I., van Renswoude, D., and Visser, I. (2020). Mixtures of Peaked Power Batschelet Distributions for Circular Data with Application to Saccade Directions. *J. Math. Psychol.* 95, 102309. doi:10.1016/j.jmp.2019.102309
- Mutasim, A. K., Stuerzlinger, W., and Batmaz, A. U. (2020). “Gaze Tracking for Eye-Hand Coordination Training Systems in Virtual Reality,” in Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–9. doi:10.1145/3334480.3382924
- Nachmias, J. (1961). Determiners of the Drift of the Eye during Monocular Fixation\*. *J. Opt. Soc. Am.* 51, 761–766. doi:10.1364/josa.51.000761
- Nakayama, M., Takahashi, K., and Shimizu, Y. (2002). “The Act of Task Difficulty and Eye-Movement Frequency for The'oculo-Motor Indices,” in Proceedings of the 2002 symposium on Eye tracking research & applications, 37–42. doi:10.1145/507072.507080
- Newhall, S. M. (1932). The Control of Eyelid Movements in Visual Experiments. *Am. J. Psychol.* 44, 555–570. doi:10.2307/1415357
- Noronha, B., Dziemian, S., Zito, G. A., Konnaris, C., and Faisal, A. A. (2017). “Wink to Grasp” - Comparing Eye, Voice & EMG Gesture Control of Grasp with Soft-Robotic Gloves,” in 2017 International Conference on Rehabilitation Robotics (ICORR) (IEEE), 1043–1048. doi:10.1109/ICORR.2017.8009387/IEEE Int. Conf. Rehabil. Robot2017
- Nyström, M., and Holmqvist, K. (2008). Semantic Override of Low-Level Features in Image Viewing—Both Initially and Overall. *J. Eye Move. Res.* 2. doi:10.16910/jemr.2.2.2
- O'Driscoll, G. A., and Callahan, B. L. (2008). Smooth Pursuit in Schizophrenia: a Meta-Analytic Review of Research since 1993. *Brain Cogn.* 68, 359–370.
- Oh, J., Han, M., Peterson, B. S., and Jeong, J. (2012). Spontaneous Eyeblinks Are Correlated with Responses during the Stroop Task. *PloS one* 7, e34871. doi:10.1371/journal.pone.0034871
- Ohtani, A. (1971). An Analysis of Eye Movements during a Visual Task. *Ergonomics* 14, 167–174. doi:10.1080/00140137108931235
- Otero-Millan, J., Troncoso, X. G., Macknik, S. L., Serrano-Pedraza, I., and Martinez-Conde, S. (2008). Saccades and Microsaccades during Visual Fixation, Exploration, and Search: Foundations for a Common Saccadic Generator. *J. Vis.* 8, 21. doi:10.1167/8.14.21
- Pan, J., Ferrer, C. C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., et al. (2017). *Salgan: Visual Saliency Prediction with Generative Adversarial Networks*. arXiv preprint arXiv:1701.01081
- Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the Role of Saliency in the Allocation of Overt Visual Attention. *Vis. Res.* 42, 107–123. doi:10.1016/s0042-6989(01)00250-4
- Patel, S. S., Jankovic, J., Hood, A. J., Jeter, C. B., and Sereno, A. B. (2012). Reflexive and Volitional Saccades: Biomarkers of huntington Disease Severity and Progression. *J. Neurol. Sci.* 313, 35–41. doi:10.1016/j.jns.2011.09.035
- Pavlović, N., and Jensen, K. (2009). Eye Tracking Translation Directionality. *Translation Res. projects* 2, 93–109.
- Peterson, J., and Allison, L. W. (1931). Controls of the Eye-Wink Mechanism. *J. Exp. Psychol.* 14, 144–154. doi:10.1037/h0070197
- Peysakhovich, V., Vachon, F., and Dehais, F. (2017). The Impact of Luminance on Tonic and Phasic Pupillary Responses to Sustained Cognitive Load. *Int. J. Psychophysiology* 112, 40–45. doi:10.1016/j.ijpsycho.2016.12.003
- Phillips, M. H., and Edelman, J. A. (2008). The Dependence of Visual Scanning Performance on Search Direction and Difficulty. *Vis. Res.* 48, 2184–2192. doi:10.1016/j.visres.2008.06.025
- Ponsoda, V., Scott, D., and Findlay, J. M. (1995). A Probability Vector and Transition Matrix Analysis of Eye Movements during Visual Search. *Acta psychologica* 88, 167–185. doi:10.1016/0001-6918(95)94012-y
- Posner, M. I., Snyder, C. R., and Davidson, B. J. (1980). Attention and the Detection of Signals. *J. Exp. Psychol. Gen.* 109, 160–174. doi:10.1037/0096-3445.109.2.160
- Pumphrey, R. J. (1948). The Theory of the Fovea. *J. Exp. Biol.* 25, 299–312. doi:10.1242/jeb.25.3.299
- Ranti, C., Jones, W., Klin, A., and Shultz, S. (2020). Blink Rate Patterns Provide a Reliable Measure of Individual Engagement with Scene Content. *Sci. Rep.* 10, 1–10. doi:10.1038/s41598-020-64999-x
- Rashbass, C. (1961). The Relationship between Saccadic and Smooth Tracking Eye Movements. *J. Physiol.* 159, 326–338. doi:10.1113/jphysiol.1961.sp006811
- Ratliff, F., and Riggs, L. A. (1950). Involuntary Motions of the Eye during Monocular Fixation. *J. Exp. Psychol.* 40, 687–701. doi:10.1037/h0057754
- Rayner, K. (1979). Eye Guidance in reading: Fixation Locations within Words. *Perception* 8, 21–30. doi:10.1068/p080021
- Rayner, K. (2012). *Eye Movements and Visual Cognition: Scene Perception and reading*. Springer Science & Business Media.
- Rayner, K. (1978). Eye Movements in reading and Information Processing. *Psychol. Bull.* 85, 618–660. doi:10.1037/0033-2909.85.3.618
- Rayner, K., Pollatsek, A., Ashby, J., and Clifton, C., Jr (2012). *Psychology of Reading*. 1 edn. Hove, East Sussex, UK: Psychology Press.
- Rayner, K. (1975). The Perceptual Span and Peripheral Cues in reading. *Cogn. Psychol.* 7, 65–81. doi:10.1016/0010-0285(75)90005-5
- Recarte, M. A., and Nunes, L. M. (2000). Effects of Verbal and Spatial-Imagery Tasks on Eye Fixations while Driving. *J. Exp. Psychol. Appl.* 6, 31–43. doi:10.1037/1076-898x.6.1.31
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). “You Only Look once: Unified, Real-Time Object Detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 779–788. doi:10.1109/cvpr.2016.91
- Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., et al. (2016). Pupil Fluctuations Track Rapid Changes in Adrenergic and Cholinergic Activity in Cortex. *Nat. Commun.* 7, 1–7. doi:10.1038/ncomms13289
- Reingold, E. M., Reichle, E. D., Glaholt, M. G., and Sheridan, H. (2012). Direct Lexical Control of Eye Movements in reading: Evidence from a Survival

- Analysis of Fixation Durations. *Cogn. Psychol.* 65, 177–206. doi:10.1016/j.cogpsych.2012.03.001
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). *Faster R-Cnn: Towards Real-Time Object Detection with Region Proposal Networks*. arXiv preprint arXiv:1506.01497.
- Renner, P., and Pfeiffer, T. (2017). “Attention Guiding Techniques Using Peripheral Vision and Eye Tracking for Feedback in Augmented-Reality-Based Assistance Systems,” in 2017 IEEE Symposium on 3D User Interfaces (3DUI) (IEEE), 186–194. doi:10.1109/3dUI.2017.7893338
- Reppert, T. R., Lempert, K. M., Glimcher, P. W., and Shadmehr, R. (2015). Modulation of Saccade Vigor during Value-Based Decision Making. *J. Neurosci.* 35, 15369–15378. doi:10.1523/jneurosci.2621-15.2015
- Rerhaye, L., Blaser, T., and Alexander, T. (2018). “Evaluation of the index of Cognitive Activity (Ica) as an Instrument to Measure Cognitive Workload under Differing Light Conditions,” in Congress of the International Ergonomics Association (Springer), 350–359. doi:10.1007/978-3-319-96059-3\_38
- Riggs, L. A., Kelly, J. P., Manning, K. A., and Moore, R. K. (1987). Blink-related Eye Movements. *Invest. Ophthalmol. Vis. Sci.* 28, 334–342.
- Riggs, L. A., Ratliff, F., Cornsweet, J. C., and Cornsweet, T. N. (1953). The Disappearance of Steadily Fixated Visual Test Objects\*. *J. Opt. Soc. Am.* 43, 495–501. doi:10.1364/josa.43.000495
- Riggs, L. A., and Ratliff, F. (1951). Visual Acuity and the normal Tremor of the Eyes. *Science* 114 (2949), 17–18. doi:10.1126/science.114.2949.17
- Robinson, D. A., Gordon, J. L., and Gordon, S. E. (1986). A Model of the Smooth Pursuit Eye Movement System. *Biol. Cybern.* 55, 43–57. doi:10.1007/bf00363977
- Robinson, D. A. (1965). The Mechanics of Human Smooth Pursuit Eye Movement. *J. Physiol.* 180, 569–591. doi:10.1113/jphysiol.1965.sp007718
- Rottach, K. G., Zivotofsky, A. Z., Das, V. E., Averbuch-Heller, L., Discenna, A. O., Poonyathalang, A., et al. (1996). Comparison of Horizontal, Vertical and diagonal Smooth Pursuit Eye Movements in normal Human Subjects. *Vis. Res.* 36, 2189–2195. doi:10.1016/0042-6989(95)00302-9
- Rubaltelli, E., Agnoli, S., and Franchin, L. (2016). Sensitivity to Affective Information and Investors’ Evaluation of Past Performance: An Eye-Tracking Study. *J. Behav. Dec. Making* 29, 295–306. doi:10.1002/bdm.1885
- Rucci, M., and Poletti, M. (2015). Control and Functions of Fixational Eye Movements. *Annu. Rev. Vis. Sci.* 1, 499–518. doi:10.1146/annurev-vision-082114-035742
- Russell, D. (2005). Consecutive and Simultaneous Interpreting. *Benjamins Translation Libr.* 63, 135–164. doi:10.1075/btl.63.10rus
- Russo, M., Thomas, M., Thorne, D., Sing, H., Redmond, D., Rowland, L., et al. (2003). Oculomotor Impairment during Chronic Partial Sleep Deprivation. *Clin. Neurophysiol.* 114, 723–736. doi:10.1016/s1388-2457(03)00008-7
- Rutherford, M. D., and Towns, A. M. (2008). Scan Path Differences and Similarities during Emotion Perception in Those with and without Autism Spectrum Disorders. *J. Autism Dev. Disord.* 38, 1371–1381. doi:10.1007/s10803-007-0525-7
- Salthouse, T. A., and Ellis, C. L. (1980). Determinants of Eye-Fixation Duration. *Am. J. Psychol.* 93, 207–234. doi:10.2307/1422228
- Salvucci, D. D., and Goldberg, J. H. (2000). “Identifying Fixations and Saccades in Eye-Tracking Protocols,” in Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, New York, NY, USA (Palm Beach Gardens, FL: Association for Computing Machinery), ETRA '00, 71–78. doi:10.1145/355017.355028
- Schiller, P. H., True, S. D., and Conway, J. L. (1980). Deficits in Eye Movements Following Frontal Eye-Field and superior Colliculus Ablations. *J. Neurophysiol.* 44, 1175–1189. doi:10.1152/jn.1980.44.6.1175
- Schoonahd, J. W., Gould, J. D., and Miller, L. A. (1973). Studies of Visual Inspection. *Ergonomics* 16, 365–379. doi:10.1080/00140137308924528
- Schwalm, M., and Rosales Jubal, E. (2017). Back to Pupillometry: How Cortical Network State Fluctuations Tracked by Pupil Dynamics Could Explain Neural Signal Variability in Human Cognitive Neuroscience. *Eneuro* 4. doi:10.1523/ENEURO.0293-16.2017
- Schwiedrzik, C. M., and Sudmann, S. S. (2020). Pupil Diameter Tracks Statistical Structure in the Environment to Increase Visual Sensitivity. *J. Neurosci.* 40, 4565–4575. doi:10.1523/jneurosci.0216-20.2020
- Seeber, K. G., and Kerzel, D. (2012). Cognitive Load in Simultaneous Interpreting: Model Meets Data. *Int. J. Bilingualism* 16, 228–242. doi:10.1177/1367006911402982
- Semmelmann, K., and Weigelt, S. (2018). Online Webcam-Based Eye Tracking in Cognitive Science: A First Look. *Behav. Res.* 50, 451–465. doi:10.3758/s13428-017-0913-7
- Sewell, W., and Komogortsev, O. (2010). “Real-time Eye Gaze Tracking with an Unmodified Commodity Webcam Employing a Neural Network,” in CHI’10 Extended Abstracts on Human Factors in Computing Systems, 3739–3744. doi:10.1145/1753846.1754048
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x
- Shiferaw, B., Downey, L., and Crewther, D. (2019). A Review of Gaze Entropy as a Measure of Visual Scanning Efficiency. *Neurosci. Biobehavioral Rev.* 96, 353–366. doi:10.1016/j.neubiorev.2018.12.007
- Shin, Y. S., Chang, W.-d., Park, J., Im, C.-H., Lee, S. I., Kim, I. Y., et al. (2015). Correlation between Inter-blink Interval and Episodic Encoding during Movie Watching. *PLoS one* 10, e0141242. doi:10.1371/journal.pone.0141242
- Simons, D. J., and Chabris, C. F. (1999). Gorillas in Our Midst: Sustained Inattention Blindness for Dynamic Events. *perception* 28, 1059–1074. doi:10.1068/p2952
- Snider, J., Spence, R. J., Engler, A.-M., Moran, R., Hacker, S., Chukoskie, L., et al. (2021). *Distraction “Hangover”: Characterization of the Delayed Return to Baseline Driving Risk after Distracting Behaviors*. Thousand Oaks, CA: Human factors, 00187208211012218.
- Spering, M., and Gegenfurtner, K. R. (2007). Contextual Effects on Smooth-Pursuit Eye Movements. *J. Neurophysiol.* 97, 1353–1367. doi:10.1152/jn.01087.2006
- Stapel, J., El Hassnaoui, M., and Happee, R. (2020). *Measuring Driver Perception: Combining Eye-Tracking and Automated Road Scene Perception*. Thousand Oaks, CA: Human factors, 00187208209595958.
- Staub, A., and Benatar, A. (2013). Individual Differences in Fixation Duration Distributions in reading. *Psychon. Bull. Rev.* 20, 1304–1311. doi:10.3758/s13423-013-0444-x
- Steinman, R. M., Haddad, G. M., Skavenski, A. A., and Wyman, D. (1973). Miniature Eye Movement. *Science* 181, 810–819. doi:10.1126/science.181.4102.810
- Stern, J. A., Walrath, L. C., and Goldstein, R. (1984). The Endogenous Eyeblink. *Psychophysiology* 21, 22–33. doi:10.1111/j.1469-8986.1984.tb02312.x
- Stevenson, S. B., Volkman, F. C., Kelly, J. P., and Riggs, L. A. (1986). Dependence of Visual Suppression on the Amplitudes of Saccades and Blinks. *Vis. Res.* 26, 1815–1824. doi:10.1016/0042-6989(86)90133-1
- Sun, F., Tauchi, P., and Stark, L. (1983). Dynamic Pupillary Response Controlled by the Pupil Size Effect. *Exp. Neurol.* 82, 313–324. doi:10.1016/0014-4886(83)90404-1
- Sweller, J. (2011). “Cognitive Load Theory,” in *Psychology of Learning and Motivation* (Elsevier), 55, 37–76. doi:10.1016/b978-0-12-387691-1.00002-8
- Takeda, K., and Funahashi, S. (2002). Prefrontal Task-Related Activity Representing Visual Cue Location or Saccade Direction in Spatial Working Memory Tasks. *J. Neurophysiol.* 87, 567–588. doi:10.1152/jn.00249.2001
- Termsarasab, P., Thammongkolchai, T., Rucker, J. C., and Frucht, S. J. (2015). The Diagnostic Value of Saccades in Movement Disorder Patients: a Practical Guide and Review. *J. Clin. Mov. Disord.* 2, 14–10. doi:10.1186/s40734-015-0025-4
- Tullis, T., and Albert, B. (2013). *Measuring the User Experience*. Elsevier, 163–186. chap. Behavioral and physiological metrics. doi:10.1016/b978-0-12-415781-1.00007-8 Behavioral and Physiological Metrics
- Valls-Sole, J. (2019). Spontaneous, Voluntary, and Reflex Blinking in Clinical Practice. *J. Clin. Neurophysiol.* 36, 415–421. doi:10.1097/wnp.0000000000000561
- Van der Stigchel, S., Rommelse, N. N. J., Deijen, J. B., Geldof, C. J. A., Witlox, J., Oosterlaan, J., et al. (2007). Oculomotor Capture in Adhd. *Cogn. Neuropsychol.* 24, 535–549. doi:10.1080/02643290701523546
- Van Orden, K. F., Jung, T.-P., and Makeig, S. (2000). Combined Eye Activity Measures Accurately Estimate Changes in Sustained Visual Task Performance. *Biol. Psychol.* 52, 221–240. doi:10.1016/s0301-0511(99)00043-5
- van Tricht, M. J., Nieman, D. H., Bour, L. J., Boeree, T., Koelman, J. H. T. M., de Haan, L., et al. (2010). Increased Saccadic Rate during Smooth Pursuit Eye Movements in Patients at Ultra High Risk for Developing a Psychosis. *Brain Cogn.* 73, 215–221. doi:10.1016/j.bandc.2010.05.005

- van Zoest, W., Donk, M., and Theeuwes, J. (2004). The Role of Stimulus-Driven and Goal-Driven Control in Saccadic Visual Selection. *J. Exp. Psychol. Hum. perception Perform.* 30, 746–759. doi:10.1037/0096-1523.30.4.749
- Vandenberg, L., Bouwmeester, S., Bocanegra, B. R., and Zwaan, R. A. (2013). Detecting Cognitive Interactions through Eye Movement Transitions. *J. Mem. Lang.* 69, 445–460. doi:10.1016/j.jml.2013.05.006
- Velichkovsky, B. B., Khromov, N., Korotin, A., Burnaev, E., and Somov, A. (2019). “Visual Fixations Duration as an Indicator of Skill Level in Esports,” in IFIP Conference on Human-Computer Interaction (Springer), 397–405. doi:10.1007/978-3-030-29381-9\_25
- Velichkovsky, B. M., Dornhoefer, S. M., Pannasch, S., and Unema, P. J. (2000). “Visual Fixations and Level of Attentional Processing,” in Proceedings of the 2000 symposium on eye tracking research & applications, 79–85. doi:10.1145/355017.355029
- Venjakob, A., Marnitz, T., Mahler, J., Sechelmann, S., and Roetting, M. (2012). “Radiologists’ Eye Gaze when reading Cranial Ct Images,” in Medical imaging 2012: Image perception, observer performance, and technology assessment (San Diego: International Society for Optics and Photonics), 8318, 83180B. doi:10.1117/12.913611
- Wade, M. G., and Jones, G. (1997). The Role of Vision and Spatial Orientation in the Maintenance of Posture. *Phys. Ther.* 77, 619–628. doi:10.1093/ptj/77.6.619
- Wainstein, G., Rojas-Libano, D., Crossley, N. A., Carrasco, X., Aboitiz, F., and Ossandón, T. (2017). Pupil Size Tracks Attentional Performance in Attention-Deficit/hyperactivity Disorder. *Sci. Rep.* 7, 8228–8229. doi:10.1038/s41598-017-08246-w
- Walker, R., McSorley, E., and Haggard, P. (2006). The Control of Saccade Trajectories: Direction of Curvature Depends on Prior Knowledge of Target Location and Saccade Latency. *Perception & Psychophysics* 68, 129–138. doi:10.3758/bf03193663
- Walker-Smith, G. J., Gale, A. G., and Findlay, J. M. (1977). Eye Movement Strategies Involved in Face Perception. *Perception* 6, 313–326. doi:10.1068/p060313
- Wang, C.-A., Brien, D. C., and Munoz, D. P. (2015). Pupil Size Reveals Preparatory Processes in the Generation of Pro-saccades and Anti-saccades. *Eur. J. Neurosci.* 41, 1102–1110. doi:10.1111/ejn.12883
- Wang, R. I., Pelfrey, B., Duchowski, A. T., and House, D. H. (2012). “Online Gaze Disparity via Bionocular Eye Tracking on Stereoscopic Displays,” in 2012 Second International Conference on 3D Imaging, Modeling, Processing (Visualization & Transmission/IEEE), 184–191. doi:10.1109/3dimpvt.2012.37
- Wang, Y., Lu, S., and Harter, D. (2021). Multi-sensor Eye-Tracking Systems and Tools for Capturing Student Attention and Understanding Engagement in Learning: A Review. *IEEE Sensors J.* 21, 22402–22413. doi:10.1109/jsen.2021.3105706
- Warren, D. E., Thurtell, M. J., Carroll, J. N., and Wall, M. (2013). Perimetric Evaluation of Saccadic Latency, Saccadic Accuracy, and Visual Threshold for Peripheral Visual Stimuli in Young Compared with Older Adults. *Invest. Ophthalmol. Vis. Sci.* 54, 5778–5787. doi:10.1167/iops.13-12032
- Wass, S. V., de Barbaro, K., and Clackson, K. (2015). Tonic and Phasic Co-variation of Peripheral Arousal Indices in Infants. *Biol. Psychol.* 111, 26–39. doi:10.1016/j.biopsycho.2015.08.006
- Watson, A. B., and Yellott, J. I. (2012). A Unified Formula for Light-Adapted Pupil Size. *J. Vis.* 12, 12. doi:10.1167/12.10.12
- Wedel, M., and Pieters, R. (2008). A Review of Eye-Tracking Research in Marketing. *Rev. marketing Res.*, 123–147. doi:10.4324/9781351550932-5
- Widdel, H. (1984). “Operational Problems in Analysing Eye Movements,” in *Advances in Psychology* (Elsevier), 22, 21–29. doi:10.1016/s0166-4115(08)61814-2
- Wierda, S. M., van Rijn, H., Taatgen, N. A., and Martens, S. (2012). Pupil Dilation Deconvolution Reveals the Dynamics of Attention at High Temporal Resolution. *Proc. Natl. Acad. Sci.* 109, 8456–8460. doi:10.1073/pnas.1201858109
- Xu-Wilson, M., Zee, D. S., and Shadmehr, R. (2009). The Intrinsic Value of Visual Information Affects Saccade Velocities. *Exp. Brain Res.* 196, 475–481. doi:10.1007/s00221-009-1879-1
- Yarbus, A. L. (1967). *Eye Movements and Vision*. Springer.
- Young, L. R. (1971). “Pursuit Eye Tracking Movements,” in *The Control of Eye Movements* (New York: Academic Press), 429–443. doi:10.1016/b978-0-12-071050-8.50019-7
- Young, L. R., and Sheena, D. (1975). Survey of Eye Movement Recording Methods. *Behav. Res. Methods Instrumentation* 7, 397–429. doi:10.3758/bf03201553
- Yu, G., Xu, B., Zhao, Y., Zhang, B., Yang, M., Kan, J. Y. Y., et al. (2016). Microsaccade Direction Reflects the Economic Value of Potential Saccade Goals and Predicts Saccade Choice. *J. Neurophysiol.* 115, 741–751. doi:10.1152/jn.00987.2015
- Zackon, D. H., and Sharpe, J. A. (1987). Smooth Pursuit in senescence: Effects of Target Acceleration and Velocity. *Acta oto-laryngologica* 104, 290–297. doi:10.3109/00016488709107331
- Zelinsky, G. J., and Sheinberg, D. L. (1997). Eye Movements during Parallel-Serial Visual Search. *J. Exp. Psychol. Hum. perception Perform.* 23, 244–262. doi:10.1037/0096-1523.23.1.244
- Zénon, A. (2017). Time-domain Analysis for Extracting Fast-Paced Pupil Responses. *Sci. Rep.* 7, 1–10. doi:10.1038/srep41484
- Zhang, X., Yuan, S.-M., Chen, M.-D., and Liu, X. (2018). A Complete System for Analysis of Video Lecture Based on Eye Tracking. *IEEE Access* 6, 49056–49066. doi:10.1109/access.2018.2865754
- Zuber, B. L., Stark, L., and Cook, G. (1965). Microsaccades and the Velocity-Amplitude Relationship for Saccadic Eye Movements. *Science* 150, 1459–1460. doi:10.1126/science.150.3702.1459

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Mahanama, Jayawardana, Rengarajan, Jayawardena, Chukoskie, Snider and Jayarathna. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Ten Questions for a Theory of Vision

Marco Gori<sup>1,2\*</sup>

<sup>1</sup>SAILab, Università di Siena, Siena, Italy, <sup>2</sup>3iA Chair Université Côte d'Azur, Nice, France

## OPEN ACCESS

### Edited by:

Marcello Pelillo,  
Ca' Foscari University of Venice, Italy

### Reviewed by:

Giuseppe Boccignone,  
University of Milan, Italy  
Kaleem Siddiqi,  
McGill University, Canada  
Fabio Cuzzolin,  
Oxford Brookes University,  
United Kingdom

### \*Correspondence:

Marco Gori  
marco.gori@unisi.it

### Specialty section:

This article was submitted to  
Computer Vision,  
a section of the journal  
Frontiers in Computer Science

**Received:** 27 April 2021

**Accepted:** 17 November 2021

**Published:** 03 March 2022

### Citation:

Gori M (2022) Ten Questions for a  
Theory of Vision.  
Front. Comput. Sci. 3:701248.  
doi: 10.3389/fcomp.2021.701248

By and large, the remarkable progress in visual object recognition in the last few years has been fueled by the availability of huge amounts of labelled data paired with powerful, bespoke computational resources. This has opened the doors to the massive use of deep learning, which has led to remarkable improvements on new challenging benchmarks. While acknowledging this point of view, in this paper I claim that the time has come to begin working towards a deeper understanding of visual computational processes that, instead of being regarded as applications of general purpose machine learning algorithms, are likely to require tailored learning schemes. A major claim of in this paper is that current approaches to object recognition lead to facing a problem that is significantly more difficult than the one offered by nature. This is because of learning algorithms that work on images in isolation, while neglecting the crucial role of temporal coherence. Starting from this remark, this paper raises ten questions concerning visual computational processes that might contribute to better solutions to a number of challenging computer vision tasks. While this paper is far from being able to provide answers to those questions, it contains some insights that might stimulate an in-depth re-thinking in object perception, while suggesting research directions in the control of object-directed action.

**Keywords:** computer vision, convolutional networks, biological plausibility, motion invariance, optical flow

## 1 INTRODUCTION

The construction of huge supervised visual data bases has significantly contributed to the spectacular performance of deep learning. However, the extreme exploitation of the truly artificial communication protocol of supervised learning has its drawbacks, including the vulnerability to adversarial attacks, which might also be tightly connected to the negligible role typically played to the temporal structure. Are not we missing something? It looks like Nature did a great job by using time to “sew all the video frames”, whereas it goes unnoticed to our eyes! At the dawn of pattern recognition, when we also began to cultivate the idea of interpreting natural video, in order to simplify the problem of dealing with a huge amount of information we removed time, the connecting thread between frames. As a consequence, all tasks of pattern recognition were turned into problems formulated on collections of images, where we only exploited spatial regularities and neglected the crucial role of temporal coherence. Interestingly, when considering the general problem of object recognition and scene interpretation, the joint role of the computational resources and the access to huge visual databases of supervised data has contributed to erect nowadays “reign of computer vision”. At a first glance this is reasonable, especially if you consider that video were traditionally heavy data sources to be played with. However, a closer look reveals that we are in fact neglecting a fundamental clue to interpret visual information, and that we have ended up facing problems where the extraction of the visual concepts is mostly based on spatial regularities. On the other hand, reigns have typically consolidated rules from which it's hard to escape. This is common in novels and real life. “The Three Princes of Serendip” is the English version of “Peregrinaggio di tre giovani figliuoli



del re di Serendippo,” published by Michele Tramezzino in Venice on 1557. These princes journeyed widely, and as they traveled they continually made discoveries, by accident and sagacity, of things they were not seeking. A couple of centuries later, in a letter of 28 January 1754 to a British envoy in Florence, the English politician and writer Horace Walpole coined a new term: serendipity, which is succinctly characterized as the art of finding something when searching for something else. Couldn't similar travels open new scenario in computer vision? Couldn't the visit to well-established scientific domains, where time is dominating the scene, open new doors to an in-depth understanding of vision? We need to stitch the frames to recompose the video using time as a thread, the same thread we had extracted to work on the images at the birth of the discipline. We need to go beyond a peaceful interlude and think of reinforcing the currently few contributions on learning theories based on video more than on images!

This paper is a travel towards the frontiers of the science of vision with special emphasis on object perception. We drive the discussion by a number of curious questions that mostly arise as one tries to interpret and disclose natural vision processes in a truly computational framework. Unfortunately, as yet, we are far away from addressing the posed questions. However, this paper takes the position that asking right questions on the discipline might themselves stimulate its progress.

## 2 CUTTING THE UMBILICAL CORD WITH PATTERN RECOGNITION

In the eighties, Satoru Watanabe wrote a seminal book (Watanabe, 1985) in which he pointed out the different facets of pattern recognition. Most of the modern work on computer vision for object perception fits with Watanabe's view of pattern recognition as statistical decision making and pattern recognition as categorization. Based on optimization schemes with billions of variables and universal approximation capabilities, the spectacular results of deep learning have elevated this view of pattern recognition to a position where it is hardly debatable. While the emphasis on a general theory of vision was already the main objective at the dawn of the discipline (Marr, 1982), its evolution has been mostly marked by significant experimental achievements. Most successful approaches seem to be the natural outcome of a very well-established tradition in pattern recognition methods working on images, which have given rise to the present emphasis on collecting big labelled image databases (e.g., Deng et al., 2009). However, in spite of these successful results, this could be the time of an in-depth rethinking of what we have been doing, especially by considering the remarkable traditions of the overall field of vision. Couldn't it be the right time to exploit the impressive literature in the field of vision to conquer a more unified view of object perception?

In the last few years, a number of studies in psychology and cognitive science have been pushing truly novel approaches to vision. In (Kingstone et al., 2010), it is pointed out that a critical problem that continues to bedevil the study of human cognition is related to the remarkably successes gained in experimental

psychology, where one is typically involved in simplifying the experimental context with the purpose to discover causal relationships. In so doing we minimize the complexity of the environment and maximize the experimental control, which is typically done also in computer vision when we face object recognition. However, one might ask whether such a simplification is really adequate and, most importantly, if it is indeed a simplification. Are we sure that treating vision as a collection of unrelated frames leads a simplification of learning visual tasks? In this paper we argue this not the case, since cognitive processes vary substantially with changes in context. When promoting the actual environmental interaction, Kingstone et al. (2010) introduce a novel research approach, called, “Cognitive Ethology”, where one opts to explore first how people behave in a truly natural situation. Once we have collected experience and evidence in the actual environment then we can move into the laboratory to test hypotheses. This strongly suggests that also machines should learn in the wild!

Other fundamental lessons come from the school of robotics for whatever involves the control of object-directed actions. In (Benjamin et al., 2011), it is pointed out that “the purpose of vision is very different when looking at a static scene to when engaging in real-world behavior.” The interplay between extracting visual information and coordinating the motor actions is a crucial issue to face for gaining an in-depth understanding of vision. One early realizes that manipulation of objects is not something that we learn from a picture; it looks like you definitely need to act yourself if you want to gain such a skill. Likewise, the perception of the objects you manipulate can nicely get a reinforcement form such as mechanical feedback. The mentioned interplay between perception and action finds an intriguing convergence in the natural processes of gaze control and, overall, on the focus of attention (Ballard, 1991). It looks like *animate vision* goes beyond passive information extraction and plays an important role in better posing most vision tasks.

The studies in computer vision might benefit significantly also from the exploration of the links with predictive coding (Rao and Ballard (1999) that have had a remarkable impact in neuroscience. In that framework one is willing to study theories of brain in which it constantly generates and updates a “mental model” of the environment. Overall, the model is supposed to generate its own predictions of sensory input and compare them to the actual sensory input. The prediction error is expected to be used to update and revise the mental model. While most of the studies in deep learning have been focused on the direct learning of object categories, there are a few contributions also in the direction of performing a sort of predictive coding by means of auto-encoding architectures (Ronneberger et al., 2015). This neural network, which is referred to as a *U-net*, is used for medical image segmentation.

## 3 DEALING WITH VIDEO INSTEAD OF IMAGES

While the processing of video has been massively investigated for a number of computer vision tasks, including tracking and action



recognition, the emphasis on methods rooted on still images has currently been dominating the state of the art in object recognition approaches. In this paper we argue that there are strong arguments to start exploring the more natural visual interaction that animals experience in their own environment for all perception tasks. The idea of shifting to video is very much related to the growing interest of *learning in the wild* that has been explored in the last few years<sup>1</sup>. The learning processes in the wild have a different nature with respect to those that are typically considered in machine learning. While ImageNet (Deng et al., 2009) is a collection of unrelated images, a video supports information only when motion is involved. In presence of still images that last for awhile, the corresponding stream of equal frames only conveys the information of a single image—apart from the duration of the interval in which the video has been kept constant. As a consequence, visual environments mostly diffuse information only when motion is involved. As time goes by, the information is only carried out by motion, which modifies one frame to the next one according to the optical flow. Once we deeply capture this fundamental feature of vision, we realize that a different theory of machine learning is needed that must be capable of naturally processing streams instead of a collection of independent images. An important ingredient for such a theory is that of emphasizing the role of the position (pixel) on which the decision is carried out and, even more, the role of time in the recognition processes doesn't seem to play a central role. It looks like we are mostly ignoring that we are in front of spatiotemporal information, whose reduction to isolated patterns might not be a natural approach especially for the complex tasks that we have been recently tackling. While there are already remarkable contributions on computer vision approaches that perform semantic labelling, most methods struggle for massive labeling that is difficult to achieve and definitely far away from natural human skills.

It is worth mentioning that pixel-based computations and segmentation have been successfully addressed in important real-world applications. In the last decades, the massive production of electronic documents, along with their printed version, has given rise to specialized software tools to extract textual information from optical data. Most optical documents, like tax forms or invoices, are characterized by a certain layout which dramatically simplifies the process of information extraction. Basically, as one recognizes the class of a document, its layout offers a significant prior on what we can expect to find in its different areas. For those documents, the segmentation process can often be given a somewhat formal description, so as most of the problems are reduced to deal with the presence of noise. Basically, the knowledge on the document layout typically offers the opportunity of providing robust solutions. The noise doesn't compromise significantly the presence of segmentation, that is in fact very well driven by the expectations provided in each pixel of the documents. These guidelines have been fueling the field of

document analysis and recognition (DAR), whose growth in the last few years has led to impressive results (Marinai et al., 2005). Unfortunately, in most real-world problems, as we move to natural images and vision, the methodology used in DAR is not really effective. The reason is that there is no longer a reliable anchor to which one can cling for segmenting the objects of a scene. While we can provide a clear description of characters and lines in optical documents, the same doesn't hold for the picture of a car which is mostly hidden by a truck during the overtaking. Humans exhibit a spectacular detection ability by simply relying on small glimpses at different scale and rotations. In no way are those cognitive processes reducible to the well-posed segmentation problems of chars and lines in optical documents. As we realize that there is a car, we can in fact provide its segmentation. Likewise, if an oracle gives us the segmented portion of a car, we can easily classify it. Interestingly, we don't really know which of the two processes is given a priority—if any. We are trapped into the chicken-egg dilemma on whether classification of objects must take place first of segmentation or vice versa. Amongst others, this issue has been massively investigated by in (Borenstein and Ullman, 2002) and pointed out in (Ullman, 1979). This intriguing dilemma might be connected with the absence of focus of attention, which necessarily leads to holistic mechanisms of information extraction. Unfortunately, while holistic mechanisms are required at a certain level of abstraction, the segmentation is a truly local process that involves low level features.

The bottom line is that most problems of computer vision are posed according to the historical evolution of the applications more than via an in-depth analysis of the underlying computational processes. While this choice has been proven to be successful in many real-world cases, stressing this research guideline might lead, on the long run, to sterile directions. Somewhat outside the mainstream of massive exploration of supervised learning, Poggio and Anselmi (Poggio and Anselmi, 2016) pointed out the crucial role of incorporating appropriate visual invariances into deep nets to go beyond the simple translation invariance that is currently characterizing convolutional networks. They propose an elegant mathematical framework on visual invariance and enlighten some intriguing neurobiological connections. Couldn't it be the case that the development of appropriate invariances might be exactly what is needed to go one step beyond?

## 4 QUESTIONS AND INSIGHTS

A good way to attack important problems is to pose the right question. To quote Tukey:

Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise.

Overall, posing appropriate questions can open a debate and solicit answers. However, the right questions cannot be easily posed since, while they need to have a big picture in mind, we

<sup>1</sup>See, e.g., <https://sites.google.com/site/wildml2017icml/>. Of course, the spirit of learning in the wild goes beyond video, but there is in fact a natural match between them

also need the identification of reasonable intermediate steps. It is often the case that while addressing little problems, inconsistencies arise that suggest the formulation of better questions. In this section, we formulate ten questions on the emergence of visual skills in nature that might contribute to the development of a new approach to computer vision that is based in processing of video instead of huge collections of images.

## 4.1 How can we Go Beyond “Intensive Supervision”?

We start with a fundamental question that reveals a commonly recognized limitation of current studies in computer vision:

*Q1: How can animals acquire visual skills without requiring “intensive supervision”?*

Everybody working in the broad area of artificial intelligence and in the field of cognitive science has been stimulated by this question. Yet, it has not been addressed as it definitely deserves! For example, it is well-known that the acquisition of the abstract notion of objects goes well beyond shape interpretation. It has been pointed out that a fundamental abstraction process arises when considering the object interactions in the environment, which does convey affordance (Gibson, 1950; Gibson, 1966; Gibson and Boston 1979). While this cognitive property is clearly of fundamental importance, we have not seen its significant exploitation in state of the art object recognition systems, yet.

The recent remarkable achievements in computer vision come from a different mechanism with respect to human vision which is based on tons of supervised examples—of the order of millions. The different environmental interactions that one can conceive in computers makes it possible to stress this artificial protocol of learning that is supported by mathematical foundations from decades and, more recently, by professional software tools. Interestingly, humans could not be exposed to such a boring interaction. Hence, strictly speaking, the mechanisms behind supervised learning are artificial and offer a space for machines to conquer visual skills that humans couldn't replicate in such a context. Of course, there is no need to be surprised by the spectacular capabilities that machines can achieve under such an artificial communication protocol, just like nobody gets surprised by the computer speed in performing multiplications.

On the other hand, because of the completely different environmental interaction, humans conquer the capability of recognizing objects just by a few supervisions. If we are interested in the scientific foundations of vision, we should aspire for an in-depth explanation of this remarkable difference between humans recognition capabilities and present computer vision technologies. Because of the expected saturation of performance of systems based on state of the art technologies, the time has already come to face the question. This has been advocated in a number of papers (see e.g., Lee et al., 2009; Ranzato et al., 2007; Goroshin et al., 2015; Tavaneai et al., 2016), and the interest in this kind of exploration is growing.

It turns out that when the posed question is analyzed more carefully, one easily comes into the conclusion that the answer must be searched in the different learning protocols in humans and computers (e.g., active learning and active vision, see Aloimonos et al., 1988). In a sense, we need to focus on the role of a continuous environmental interaction during the “life of the agent.” The interaction is not only carried out through symbolic communication but also it seems to rely primarily on continuous-based information exchange. The reward for successful actions is dominating the visual learning process in animals of any specie. These actions range from drinking milk from the mother's breast to obstacle avoidance. Hence, those actions basically result in reinforcement learning processes that contribute to the development of the visual skills. An eagle, like any predator, is early driven by the objective of capturing the prey, that strongly contributes to the development of the visual skills. It is worth mentioning that for newborns, like for other primates, the acquisition of visual skills also benefits from the joint development of motion control. As soon as a baby begins the first experiences of object manipulation, such a task is paired with visual development. Successful manipulations are due to the correct visual interpretation which, in turn, is reinforced by the concrete acts of touching and moving. Hence, this is just a way to “supervise” and reinforce visual skills as the outcome of other processes taking place in the environment. Interestingly, at that time, linguistic skills are nearly absent, which clearly indicates that, just for other animals, the learning of vision undergoes its own developmental phases, while the interplay with language comes later on.

Of course, in nature the discussion on the acquisition of vision cannot neglect the fundamental role of genetic inheritance, that has a fundamental impact on the cognitive developmental steps. It is worth mentioning that remarkable differences have been discovered between different species of animals, in terms of the balance between their genome and their learning acquisition from the environment. For example, while chicks acquire significant visual skills early on (Wood and Wood, 2020), it takes months for humans, that clearly need to conquer more sophisticated skills at the perceptual level.

Interestingly there are intriguing connections with recent achievements in machine learning, where we have been constantly underlying the crucial role of transfer learning mechanisms (Pan and Yang, 2010) in many real-world problems. Clearly, whenever we transfer knowledge to the learning agent, this simplifies its own visual tasks and, as a consequence, the agent can learn by relying on less supervised examples. However, one should bear in mind that the current generation of neural networks that are currently used in transfer learning relies on the development of “individuals” that are strongly limited from a “genetic viewpoint”, since they haven't been exposed to visual natural interactions. The learned features that are “genetically” transmitted are typically developed under supervised learning on classification that, even for large databases, might be biased by the specific benchmark.

**BOX 1** | The bottom line is that for the posed question to be addressed one must dramatically change the agent environmental interactions. Interestingly, it is not the lack of technology for mimicking human interaction that mostly prevents us from going beyond “intensive supervision”, but the lack of a theory to support the appropriate computational mechanisms. From one side, this is a stimulating scientific challenge. From the other side, the development of a similar theory would likely contribute to open new technological perspectives where machines learn to see by a on-line scheme without needing visual databases.

## 4.2 What is the Role of Time?

As stated at the end of the previous section, a remarkable distinction between humans and state of the art machines is that they operate within a very different learning environment. The visual interaction that we experiment in nature leads us to face the following fundamental question:

*Q2: How can animals gradually acquire visual skills in their own environments?*

Apparently, such a gradual learning scheme also takes place in machine learning. But, can we really state that the “gradual” process of human learning is somewhat related to the “gradual” weight updating of neural networks? A closer look at the mechanisms that drive learning in vision tasks suggests that current models of machine learning mostly disregard the fundamental role of “time”, which shouldn’t be confused with the iteration steps that mark the weight update. First, notice that learning to see in nature takes place in a context where the classic partition into learning and test environment is arguable. On the other hand, this can be traced back to early ideas on statistical machine learning and pattern recognition, which are dominated by the principles of statistics. In the extreme case of batch mode learning, the protocol assumes that the agent possesses information on its life before coming to life. Apparently this doesn’t surprise computer vision researchers, whereas it sounds odd for the layman, whose viewpoint shouldn’t be neglected, since we might be trapped into an artificial world created when no alternative choice was on the horizon. The adoption of mini-batches and even the extreme solution of on-line stochastic gradient learning are still missing a true incorporation of time. Basically, they pass through the whole training data many times, a process which is still far from natural visual processes, where the causal structure of time dictates the gradual exposition to video sources. There is a notion of life-long learning that is not captured in current computational schemes, since we are ignoring the role of time which imposes causality.

Interestingly, when we start exploring the alternatives to huge collections of labelled images, we are immediately faced with a fundamental choice, which arises when considering their replacement with video collections. What about their effectiveness? Suppose you want to prepare a collection to be used for learning the market segment associated with cars (luxury vehicles, sport cars, SUVs/off-road vehicles, . . . ). It could be the case that a relatively small image database composed of a few

thousands of labelled examples is sufficient to learn the concept. On the other hand, in a video setting, this corresponds with a few minutes of video, a time interval in which it is unreasonable to cover the variety of car features that can be extracted from 10,000 images! Basically, there will be a lot of near-repetitions of frames which support scarce information with respect to the abrupt change from picture to picture. This is what motivates a true paradigm shift in the formulation of learning for vision. In nature, it is unlikely to expect the emergence of vision from the accumulation of video. Hence, couldn’t we do same? A new communication protocol can be defined where the agent is simply expected to learn by processing the video as time goes by without its recording and by handling human-like vocal interactions. Interestingly, this gets rid of the need to accumulate and properly handle huge collections of labelled images, which would represent a paradigm shift in computer vision and might opens great opportunities to all research centers to compete in another battlefield.

At any stage of child development, it looks like only the visual skills that are required to face the current tasks are acquired. One might believe that this is restricted to natural processes, but we conjecture that the temporal dimension plays a crucial role in the well-positioning of most challenging cognitive tasks, regardless of whether they are faced by humans or machines. The formulation of learning in the temporal dimension likely becomes more and more important when we begin to address a number of challenges that are also outlined in the other questions posed in the paper. The role of time becomes crucial when considering the extraction of good features. This is in fact an issue that, as the interest in transfer learning has been demonstrating, has becoming more and more relevant. While in the literature we have been typically concerned with feature extraction that is independent of classic geometric transformation, it looks like we are still missing the astonishing human skill of capturing distinctive features when looking at ironed and rumpled shirts! There is no apparent difficulty to recognize shirts by keeping the recognition coherence in case we roll up the sleeves, or we simply curl them up into a ball for the laundry basket. Of course, there are neither rigid transformations, like translations and rotation, nor scale maps that transforms an ironed shirt into the same shirt thrown into the laundry basket. Is there any natural invariance?

In this paper, we claim that motion invariance is in fact the only one that we need. Translation, scale, and rotation invariances, that have been the subject of many studies, are in fact instances of invariances that can be fully gained whenever we develop the ability to detect features that are invariant under motion. If my finger moves closer and closer to my eyes then any of its representing features that is motion invariant will also be scale invariant. The finger will become bigger and bigger as it approaches my face, but it is still my finger! Clearly, translation, rotation, and complex deformation invariances derive from motion invariance. Humans life always experiments motion, so as the gained visual invariances naturally arise from motion. Studies on different types of invariances in vision have been so rich and massively investigated that one can suspect that there is something missing in the claim that enforcing motion invariance only suffices to learn. The

emergence of information-based laws of learning can take place and, consequently, a natural formulation of learning to see in the temporal dimension relies on the principle of *Material Point Invariance*, which blesses the pairing of any feature of a given object, including the brightness, with its own velocity.

We can somewhat parallel the idea of brightness invariance that is used for the estimation of the optical flow for imposing a fundamental invariance condition on any visual feature  $\varphi$ . Hence, the following consistency condition holds true:

$$\varphi(x_\varphi(t), t) = \varphi(x_\varphi(0), 0) = c_\varphi, \quad \forall t \in [0, T] \quad (1)$$

where  $x_\varphi(t)$  denotes a trajectory of the associated feature  $\varphi$  and  $c_\varphi \in \mathbb{R}$ . It is worth mentioning that motion invariance is not always desirable since perception and action also need to develop features that provide different reactions in front of motion. For example, this makes it possible to learn the meaning of “vertical” and “horizontal” positions and to react to moving objects. When thinking of the trajectory  $x_\varphi(t)$  we must bear in mind that if we consider that feature extraction takes place under focus of attention mechanisms then the visual agent always experiences motion. Formally, for any pair  $(x, t)$  (pixel-time), let  $x_\varphi(t) = x$  and  $\dot{x}_\varphi(t) = v_\varphi(x, t)$  be. Given the optical flow  $v_\varphi$ . Then motion invariance of feature  $\varphi$  can be expressed by

$$\frac{d\varphi(x_\varphi(t), t)}{dt} = \nabla \varphi(x, t) \cdot v_\varphi(x, t) + \varphi_t(x, t) = 0. \quad (2)$$

This shares the formal structure of the transport equation of brightness invariance, but it is important to notice that this equation is *stating a principle which is not supposed to be violated*. While brightness invariance represents an approximation, the conjunction stated by **Eq. 2** is expected to hold perfectly. Basically  $(\varphi, v_\varphi)$  is an *indissoluble pair* that plays a fundamental role in the learning of the visual features that characterizes the object. Notice that, just like the color components of a color video typically share the same velocity, also different features can be aggregated with the same velocity.

What if an object is gradually deformed into another one? One can think of cat which is very slowly transformed to a dog! This relates to the metaphysical question of whether and how things persist over time. Philosophers regards things as concrete material objects as well as pure abstract objects which are associated with concepts and ideas<sup>2</sup>. Concepts can drift as video changes very slowly, which can deceive any intelligent agent who relies on motion invariance only, provided that such an agent fails at detecting slow motion. While one can always argue about the possibility of achieving certain thresholds for slow motion detection, when trusting motion invariance one must be ready to accept the problem of concept drift. Interestingly, as will be pointed out in the reminder of the

paper, the development of focus of attention mechanisms helps facing also this problem.

It could be the case that the extraction of very efficient information from visual processes that has captured the attention for decades of computer scientists and engineers benefits from a sort of *pre-algorithmic phase* where we primarily need to understand basic *perceptual laws of vision* that hold regardless of the nature of the agent. This research guideline has been stimulating the conception of variational laws of learning that can capture the elegance and the simplicity of natural behavior (Betti and Gori, 2016; Betti et al., 2018). When following this approach one promotes the role of time by following principles that have a very well-established tradition in physics. Most importantly, one begins to challenge the indisputable principle that learning can be regarded as the outcome of an optimization process that operates on the risk function. Clearly, there is something wrong with this principle in nature, since the agent doesn't possess the risk at the time of its birth! While such a risk can be gradually constructed and the adoption of stochastic gradient is a powerful idea for capturing the underlying statistics, we are basically attacking a different problem with respect to what all species of animals are expected to face in nature.

**BOX 2 |** The bottom line is that while we struggle for the acquisition of huge labeled databases, the true incorporation of time might led to a paradigm shift in the process of feature extraction. We promote the study of the agent life based on the ordinary notion of time, which emerges in all its facets. The incorporation of motion invariance might be the key for overcoming the artificial protocol of supervised learning. We claim that such an invariance is in fact the only one that we need.

### 4.3 Can Animals see in a World of Shuffled Frames?

One might figure out what human life could have been in a world of visual information with shuffled frames.

**Q3:** *Could children really acquire visual skills in such an artificial world, which is the one we are presenting to machines? Don't shuffled visual frames increase the complexity of learning to see?*

A related issue has been faced in (Wood, 2016) for the acquisition of visual skills in chicks. It is pointed out that “when newborn chicks were raised with virtual objects that moved smoothly over time, the chicks developed accurate color recognition, shape recognition, and color-shape binding abilities.” Interestingly, the authors notice that in contrast, “when newborn chicks were raised with virtual objects that moved non-smoothly over time, the chicks’ object recognition abilities were severely impaired.” When exposed to a video composed of independent frames taken from a visual database, like ImageNet, that are presented at classic cinema frame rate of 24 fps, humans seem to experiment related difficulties in non-smooth visual presentation.

<sup>2</sup>This discussion was stimulated by Marcello Pelillo who pointed out intriguing links with The “Ship of Theseus Puzzle” and “The Puzzle of the Statue and the Clay.”



Hence, it turns out that our spectacular visual skills completely collapse in a task that is successfully faced in computer vision! As a consequence, one might start formulating conjectures on the inherent difficulty of artificial versus natural visual tasks. The remarkably different performance of humans and machines has stimulated the curiosity of many researchers in the field. Of course, you can start noticing that in a world of shuffled frames, a video requires an order of magnitude more information for its storing than the corresponding temporally coherent visual stream. This is a serious warning that is typically neglected in computer vision, since it suggests that any recognition process is likely to be more difficult when shuffling frames. One needs to extract information by only exploiting spatial regularities in the retina, while disregarding the spatiotemporal structure that is offered by nature. The removal of the thread that nature used to sew the visual frames might prevent us from the construction of a good theory of vision. Basically, we need to go beyond the current scientific peaceful interlude and abandon the safe model of restricting computer vision to the processing of images. Working with video was discouraged at the dawn of computer vision because of the heavy computational resources that it requires, but the time has come to reconsider significantly this implicit choice. Not only it is the case that humans and animals cannot see in a world of shuffled frames, but it is likely that they could not learn to see in such an environment. Shuffling visual frames is the implicit assumption of most of present vision technology that, as stated in the previous section, corresponds with neglecting the role of time in the discovery of visual regularities. No matter what computational scheme we conceive, the presentation of frames where we have removed the temporal structure exposes visual agents to a problem where a remarkable amount of information is delivered at any presentation of new examples. When going back to the previous discussion on time, one clearly see its natural environmental flow that must be somehow synchronized with the agent's computational capability. The need for this synchronization is in fact one of the reasons for focussing attention at specific positions in the retina, which confers the agent also the gradual capability of extracting information at pixel label. Moreover, as already pointed out, we need to abandon the idea of recording a data base for statistical assessment. There is nothing better than human evaluation in perceptual tasks, which could stimulate new ways of measuring the scientific progress of the discipline (see **Section 5**).

The reason for formulating a theory of learning on video instead of on images is not only rooted in the curiosity of grasping the computational mechanisms that take place in nature. A major claim in this paper is that those computational mechanisms are also fundamental in most of computer vision tasks.

**BOX 3** | It appears that, while ignoring the crucial role of temporal coherence, the formulation of most of present computer vision tasks lead us to tackle problems that are remarkably more difficult than those nature has prepared for us!

## 4.4 How can Humans Perform Pixel Semantic Labeling?

Many object recognition systems are based on an opportune pre-processing of video information represented by a vector, which is subsequently processed for class prediction. Surprisingly enough, the state of the art approaches that follow this guideline already offer quite accurate performance in real-world contexts, without relying on the semantic labelling of each pixel. Basically, a global computational scheme emerges that is typically made more and more effective when the environment in which the machine is supposed to work is quite limited, and it is known in advance. The number of the classes that one expects to recognize in the environment affects the performance, but very high accuracy can be achieved without necessarily being able to perform the object segmentation and, therefore, without needing to perform pixel semantic labeling. However, for an agent to conquer visual capabilities in a broad context, it seems to be very useful to rely on more specific visual skills. When thinking of a video, the information that one can extract is not only driven by time but also by spatial information. We humans can easily describe a scene by locating the objects in specific positions, and we can describe their eventual movement. This requires a deep integration of visual and linguistic skills, that are required to come up with compact, yet effective video descriptions. However, in any case humans can successfully provide a very accurate labeling of single pixels, which leads us to pose the following question:

**Q4:** *How can humans exhibit such an impressive skill of properly labelling single pixels without having received explicit pixel-wise supervisions? Is it not the case that such a skill must be a sort of “visual primitive” that cannot be ignored for efficiently conquering additional skills on object recognition and scene interpretation?*

Interestingly, in humans semantic pixel labelling is by driven by the focus of attention, another fundamental features that, as we will see in the remainder of the paper, is at the core of all important computational processes of vision. While pixel-based decisions are inherently interwound with a certain degree of ambiguity, they are remarkably effective. The linguistic attributes that we can extract are related to the context of the pixel that is taken into account for label attachment, while the ambiguity is mostly a linguistic more than a visual issue. In a sense, this primitive is likely in place for conquering higher abstraction levels. How can this be done? The focus on single pixels allows us to go beyond object segmentation based on sliding windows. Instead of dealing with object proposals (Zitnick and Dollár, 2014), a more primitive task is that of attaching symbols to single pixels in the retina. The task of semantic pixel labelling leads to focussing attention on the given pixel, while considering the information in its neighborhood. This clearly opens the doors to an in-depth re-thinking of pattern recognition processes. It is not only the frame content, but also where we focus attention in the retina that does matter.

Human ability of exhibiting semantic labeling at pixel level is really challenging. The visual developmental processes conquer this ability nearly without pixel-based supervisions. It seems that such a skill is mostly the outcome of the acquisition of the capability to perform object segmentation. This is obtained by constructing the appropriate memberships of the pixels that define the segmented regions. When thinking of the classic human communication protocols, one early realizes that even though it is rare to provide pixel-based supervision, the information that is linguistically conveyed to describe visual scenes makes implicit reference to the focus of attention. This holds regardless of the scale of the visual entity being described. Hence, the emergence of the capability of performing pixel semantic label seems to be deeply related to the emergence of focus of attention mechanisms. The most striking question, however, is how can humans construct such a spectacular segmentation without a specific pixel-based supervision! Interestingly, we can focus on a pixel and attach meaningful labels, without having been instructed for that task.

**BOX 4 |** The primitive of pixel semantic labelling is likely crucial for the construction of human-like visual skills. There should be a hidden supervisor in nature that, so far, has nearly been neglected. We conjecture that it is the optical flow which plays the central role for object recognition. The decision on its recognition must be invariant under motion, a property that does require a formulation in the temporal direction.

## 4.5 What is the Role of Receptive Fields and Hierarchical Architectures?

Beginning from early studies on the visual structure of the cortex (Hubel and Wiesel, 1962), neuroscientists have gradually gained evidence that it presents a hierarchical structure and that neurons process the video information on the basis of inputs restricted to receptive fields. Interestingly, the recent spectacular results of convolutional neural networks suggests that hierarchical structures based on neural computation with receptive fields play a fundamental role also in artificial neural networks (LeCun et al., 2015). The following questions naturally arise:

**Q5:** *Why are the visual mainstreams organized according to a hierarchical architecture with receptive fields? Is there any reason why this solution has been developed in biology? Why is its “replication” in neural networks so successful?*

First of all, we can promptly realize that, even though neurons are restricted to compute over receptive fields, deep structures rely on large virtual contexts for their decision. As we increase the depth of the neural network, the consequent pyramidal dependence that is established by the receptive fields increases the virtual input window used for the decision, so that higher abstraction is progressively gained as we move towards the output. Hence, while one gives up to exploit all the information available at a certain layer, the restriction to receptive field does not prevent from considering large

windows for the decision. The marriage of receptive fields with deep nets turns out to be an important ingredient for a parsimonious and efficient implementation of both biological and artificial networks. In convolutional neural networks, the assumption of using receptive fields comes with the related hypothesis of *weight sharing* on units that are supposed to extract the same feature, regardless of where the neurons are centered in the retina. In so doing we enforce the extraction of the same features across the retina. This makes sense whereas, in general, it does not make sense to extract features depending on the pixel in the retina. The same visual clues are clearly positioned everywhere in the retina and the equality constraints on the weights turn out to be a precise statement for implementing a sort of *invariance under translation*.

Clearly, this constraint has neither effect on invariance under scale nor under rotation. Any other form of invariance that is connected with deformable objects is clearly missed and is supposed to be learned. The current technology of convolutional neural networks in computer vision typically gains these invariances thanks to the power of supervised learning by “brute force.” Notice that since most of the tasks involve object recognition in a certain environment, the associated limited amount of visual information allows to go beyond the principle of extracting visual features at pixel level. Visual features can be shared over small windows in the retina by the process of pooling, thus limiting the dimension of the network. Basically, the number of features to be involved has to be simply related to the task at hand, and we can go beyond the association of the features with the pixels. However, the acquisition of human-like visual skills is not compatible with this kind of simplifications since, as stated in the previous section, humans can perform pixel semantic labeling. There is a corresponding trend in computer vision where convolutional nets are designed to keep the connection with each pixel in the retina at any layer so as to carry out segmentation and semantic pixel label. Interestingly, this is where we need to face a grand challenge. So far, very good results have been made possible by relying on massive labelling of collections of images. While image labeling for object classification is a boring task, human pixel labeling (segmentation) is even worse! Instead of massive supervised labelling, one could realize that motion and focus of attention can be massively exploited to learn the visual features mostly in an unsupervised way. A recent study in this direction is given in (Betti and Gori, 2018), where the authors provide evidence of the fact that receptive fields do favor the acquisition of motion invariance which, as already stated, is the fundamental invariance of vision. The study of motion invariance leads to dispute the effectiveness and the biological plausibility of convolutional networks. First, while weight sharing is directly gained by translational invariance on any neuron, the vice versa clearly does not hold. Hence, we can think of receptive field based neurons organized in a hierarchical architecture that carry out translation invariance without sharing their weights. This is strongly motivated also by the arguable biological plausibility of the mechanism of weight sharing (Ott et al., 2020). Such a lack of plausibility is more serious than the supposed lack of a local computational scheme in Backpropagation, which mostly comes

from the lack of delay in the forward model of the neurons (Betti and Gori, 2019).

**BOX 5 |** Hierarchical architectures and receptive fields seems to be tightly connected in the development of abstract representations. However, we have reasons to doubt that weight sharing happens in biological networks and that the removal of this constraint facilitates the implementation of motion invariance. The architectural incorporation of this fundamental invariance property, as well as the match with the need for implementing the focus of attention mechanisms likely needs neural architectures that are more sophisticated than current convolutional neural networks. In particular, neurons which provide motion invariance likely benefit from dropping the weight sharing constraint.

## 4.6 Why Two Different Main Visual Processing Streams?

In **Section 2** we have emphasized the importance of bearing in mind the neat functional distinction between vision for action and vision for perception. A number of studies in neuroscience lead to the conclusion that the visual cortex of humans and other primates is composed of two main information pathways that are referred to as the ventral stream and the dorsal stream (Goodale and Milner, 1992; Goodale and Keith Humphrey, 1998). We typically refer to the ventral “what” and the dorsal “where/how” visual pathways. The ventral stream is devoted to perceptual analysis of the visual input, such as object recognition, whereas the dorsal stream is concerned with providing spatial localization and motion ability in the interaction with the environment. The ventral stream has strong connections to the medial temporal lobe (which stores long-term memories), the limbic system (which controls emotions), and the dorsal stream. The dorsal stream stretches from the primary visual cortex (V1) in the occipital lobe forward into the parietal lobe. It is interconnected with the parallel ventral stream which runs downward from V1 into the temporal lobe.

**Q6:** *Why are there two different mainstreams? What are the reasons for these a different neural evolutions?*

This neurobiological distinction arises for effectively facing visual tasks that are very different. The exhibition of perceptual analysis and object recognition clearly requires computational mechanisms that are different with respect to those required for estimating the scale and the spatial position. Object recognition requires the ability of developing strong invariant properties that mostly characterize the objects themselves. By and large scientists agree that objects must be recognized independently of their position in the retina, scale, and orientation. While we subscribe to this point of view, a more careful analysis of our perceptual capabilities indicates that these desirable features are likely more adequate to understand the computational mechanisms behind the perception of rigid objects. The elastic deformation and the projection into the retina gives in fact rise to remarkably more complex patterns that can hardly be interpreted in the framework of geometrical invariances. We reinforce the claim that *motion*

*invariance is in fact the only invariance which does matter.* Related studies in this direction can be found (Bertasio et al., 2021). As the nose of a teddy bear approaches child’s eyes it becomes larger and larger. Hence, scale invariance is just a byproduct of motion invariance. The same holds true for rotation invariance. Interestingly, as a child deforms the teddy bear a new visual pattern is created that, in any case, is the outcome of the motion of “single object particles.” The neural enforcement of motion invariance likely takes place in the “what” neurons. Of course, neurons with built-in motion invariance are not adequate to make spatial estimations or detection of scale/rotation. Unlike the “what” neurons, in this case motion does matter and the neural response must be affected by the movement.

**BOX 6 |** These analyses are consistent with neuroanatomical evidence and suggest that “what” and “where” neurons are important also in machines. The anatomical difference between the two processing streams is in fact the outcome of a different functional role. While one can ignore such a difference and rely on the rich representational power of big deep networks, the underlined difference stimulates the curiosity of discovering canonical neural structures to naturally incorporate motion invariance, with the final purpose being that of discovering different features for perception and action.

## 4.7 Why do Some Animals Focus Attention?

It is well-known that the presence of the fovea in the retina leads to focus attention on details in the scene. Such a specialization of the visual system is widespread among vertebrates, it is present in some snakes and fishes, but among mammals is restricted to haplorhine primates. In some nocturnal primates, like the owl monkey and in the tarsier, the fovea is morphologically distinct and appears to be degenerate. An owl monkey’s visual system is somewhat different from other monkeys and apes. As its retina develops, its dearth of cones and its surplus of rods mean that this focal point never forms. Basically, a fovea is most often found in diurnal animals, thus supporting the idea that it is supposed to play an important role for capturing details of the scene (Ross, 2004). But why haven’t many mammals developed such a rich vision system based on foveate retinas? Early mammals, which emerged in the shadow of the dinosaurs, were likely forced to live nocturnal lives so as to avoid to become their prey (Sohn, 2019). In his seminal monograph, Gordon Lynn Walls (Walls, 1942) proposed that there has been a long nocturnal evolution of mammals’ eyes, which is the reason of the remarkable differences with respect to those of other vertebrates. The idea became known as the “nocturnal bottleneck” hypothesis (Gerkema et al., 2013). Mammals’ eyes tended to resemble those of nocturnal birds and lizards, but this does not hold for humans and closely related monkeys and apes. It looks they re-evolved features useful for diurnal living after they abandoned a nocturnal lifestyle upon dinosaur extinction. It is worth mentioning that haplorhine primates are not the only mammals which focus attention in the visual environment. Most mammals have quite a well-developed visual system for dealing with details. For example, it has been shown that dogs possess quite a good visual system that share many features with

those of haplorhine primates (Beltran et al., 2014). A retinal region with a primate fovea-like cone photoreceptor density has been identified but without the excavation of the inner retina. A similar anatomical structure, that has been observed in rare human subjects, has been named fovea-plana. Basically, the results in (Beltran et al., 2014) challenge the dogma that within the phylogenetic tree of mammals, haplorhine primates with a fovea are the sole lineage in which the retina has a central bouquet of cones. In non-primate mammals, there is a central region of specialization, called the *area centralis*, which also is often located temporal to the optic axis and demonstrates a local increase in photoreceptor and retinal ganglion cell density that plays a role somehow dual with respect to the fovea. Like in haplorhine primates, in those non-primate mammals we experience focus of attention mechanisms that are definitely important from a functional viewpoint.

This discussion suggests that the evolution of animals' visual system has followed many different paths that, however, are related to focus of attention mechanisms, that are typically more effective for diurnal animals. There is, however, an evolution path which is definitely set apart, in which the frog is most classic representer. More than 60 years ago, the visual behavior of the frog posed an interesting question (Lettvin et al., 1959) which is mostly still on the table. In the words of the authors:

The frog does not seem to see or, at any rate, is not concerned with the detail of stationary parts of the world around him. He will starve to death surrounded by food if it is not moving. His choice of food is determined only by size and movement.

No mammal experiments such a surprising behavior! However, the frog is not expected to eat like mammals. When tadpoles hatch and get free, they attach themselves to plants in the water such as grass weeds, and cattails. They stay there for a few days and eat tiny bits of algae. Then the tadpoles release themselves from the plants and begin to swim freely, searching out algae, plants and insects to feed upon. At that time their visual system is ready. Their food requirements are definitely different from what mammals need and their visual system has evolved accordingly for catching flying insects. Interestingly, unlike mammals, the studies in (Lettvin et al., 1959) already pointed out that the frogs' retina is characterized by uniformly distributed receptors with neither fovea nor *area centralis*. Interestingly, this means that the frog does not focus attention by eye movements. When the discussion focuses on functional issues the following natural questions arise:

**Q7: Why are the fovea and the area centralis convenient? Why do primates and other animals focus attention, whereas others, like the frog, do not?**

One can easily argue that any action that animals carry out needs to prioritize the frontal view. On the other hand, this leads to the detriment of the peripheral vision, that is also very important. In addition, this could apply for the dorsal system

whose neurons are expected to provide information that is useful to support movements and actions. Apparently, the ventral mainstream, with neurons involved in the "what" function, does not seem to benefit from foveate eyes. Apart from recent developments, most state of the art computer vision models for object recognition, just like frogs, do not focus attention, since they carry out a uniform massive parallel computation on the retina. Just like frogs, the cameras used in computer vision applications are uniformly distributed, but machines seem to conquer human-like recognition capabilities on still images. Interestingly, unlike frogs, machines recognize quite well food properly served in a bowl. This capability might be due to the current strongly artificial communication protocol. Machines benefit from supervised learning of tons of supervised pairs, a process which, as already pointed out, cannot be sustained in nature. On the other hand, as already pointed out, in order to attack the task of understanding what is located in a certain position, it is natural to think of eyes based on fovea or on area centralis. The eye movements with the corresponding trajectory of the focus of attention (FOA) is also clearly interwound with the temporal structure of video sources. In particular, humans experience eye movements when looking at fixed objects, which means that they continually experience motion. Hence, also in case of fixed images, conjugate, vergence, saccadic, smooth pursuit, and vestibulo-ocular movements lead to the acquisition of visual information from relative motion. We claim that the production of such a continuous visual stream naturally drives feature extraction, since the corresponding convolutional filters, charged with representing features for object recognition, are expected not to change during motion. The enforcement of this consistency condition creates a mine of visual data during animal life! Interestingly, the same can happen for machines. Of course, we need to compute the optical flow at the pixel level so as to enforce the consistency of all the extracted features. Early studies on this problem (see Horn and Schunck 1981), along with related improvements (see e.g., Baker et al., 2011) suggests to determine the velocity field by enforcing brightness invariance. As the optical flow is gained, it can be used to enforce motion consistency on the visual features. These features can be conveniently combined with those responsible of representing objects. Early studies driven by these ideas are reported in Gori et al., (2016), where the authors propose the extraction of visual features as a constraint satisfaction problem, mostly based on information-theoretic principles and early ideas on motion invariance.

The following remarks on focus of attention coming from nature seem to be important for conquering efficient visual skills for any intelligent agent. Basically, it looks like we are faced with functional issues which mostly obeys information-based principles.

- *The FOA drives the definition of visual primitives at pixel level.* The already mentioned visual skill that humans possess to perform pixel semantic labeling clearly indicates the capability of focusing on specific points in the retina with high resolution. Hence, FOA is needed if we want to perform such a task.



- *Eye movements and FOA help estimating the probability distribution on the retina.* At any time a visual agent clearly needs to possess a good estimation of the probability distribution over the pixels of the retina. This is important whenever we consider visual tasks for which the position does matter. This involves both the *where* and *what* neurons. In both cases it is quite obvious that any functional risk associated with the given task should avoid reporting errors in regions of the retina where there is a uniform color. The probability distribution is of fundamental importance and it is definitely related to saliency maps built on focus of attention trajectories.
- *FOA very well fits the need for receptive fields and deep nets.* We have already discussed the marriage between receptive fields and deep networks. Interestingly, the FOA mechanisms emphasize the role of a single receptive field in the computational process that takes place at any time. The saccadic movements contribute to perform “temporally segmented computations” over the retina on the different sequences produced by micro-saccadic movements. In addition, in (Betti and Gori, 2018), the authors provide evidence of the fact that receptive fields do favor the development of biologically-plausible models based on local differential equations.
- *Eye movements and FOA are the basis for establishing invariant laws.* The interplay between the FOA and the invariance properties is the key for understanding human vision and general principles that drive object recognition and scene interpretation. In order to understand the nice circle that is established during the processes of learning in vision, let us start exploring the very nature of eye movements in humans. Basically, they produce visual sequences that are separated by saccadic movement, during which no information is acquired<sup>3</sup>. Interestingly, each of those sequences is composed of pixels that somehow share common visual features. In case of micro-saccades the corresponding micro-movements explore regions with a remarkable amount of details that are somehow characterized by certain features. The same holds true for smooth pursuit, where the invariance of the extracted feature turns out to be a sort of primitive consistency property: objects do not change during their motion. Hence, any visual feature associated with “what neurons” must be invariant under any eye movement, apart from saccadic movements. Clearly, such an invariance has a true unsupervised nature. A deep net based on the discussed convolutional structure can in fact learn a set of latent features to be motion invariant. This results in an impressive collection of “labelled data” that nature offers for free. The eye movements and the FOA significantly contribute to enhance the motion invariance since, as already pointed out, humans always experience motion in

a frame of reference located on the retina. When thinking of the “what” neurons for which invariances need to be imposed, we can promptly realize that those which are close to the output are better suited for the enforcing of invariances. It is in fact quite obvious that for many of those invariances to take place we need a strong computational capability of the “what neuron.” Interestingly, this seems to suggest that “where” neurons could better be located in the early layers of the hierarchy whereas “what” neurons require higher abstraction.

- *FOA helps disambiguation at learning time.* A puzzle is offered at learning time when two or more instance of the same object are present in the same frame, maybe with different poses and scales. The FOA in this case helps disambiguating the enforcement of motion invariance. While the enforcement of weight sharing is ideal for directly implementing translation invariance, such a constraint doesn’t facilitate other more complex invariances that can better be achieved by its removal.
- *FOA drives the temporal interpretation of scene understanding.* The importance of FOA is not restricted to feature and object invariance, since it involves also the interpretation of visual scenes. It is in fact the way FOA is driven which sequentially selects the information for conquering the scene interpretation. Depending on the purpose of the agent and on its level of scene understanding the FOA is consequently moved. This process clearly shows the fundamental role of the selection of the points where to focus attention, an issue which is described in the following section.
- *FOA helps disambiguating illusions.* Depending on where an agent with foveate eyes focuses attention, concepts that, strictly speaking, don’t exist can emerge, thus creating an illusion. A noticeable example is the Kanizsa’s triangle, but it looks like other illusions arise for related reasons. You can easily experiment that as you approach any detail, it is perfectly perceived without any ambiguity. A completion mechanism arises that leads us to perceive the triangle as soon as you move away from figure and the mechanism is favored by focussing attention on the barycenter. Interestingly, the different views coming from different points where an agent with foveate eyes focuses attention likely helps disambiguating illusions, a topic that has been recently studied in classic convolutional networks Kim et al. (2019), Baker et al. (2018).
- *FOA helps to address the problem of “concept drift”.* When discussing motion invariance we mentioned the problem of concept drift. Clearly, this could dramatically affect the practical implementation of the motion invariance. However, amongst different types of FOA trajectories, the saccadic movements play the fundamental role of resetting the process, which clearly faces directly problems of concept drift.

<sup>3</sup>There is in fact a rich literature on this topic, from which it is clearly stated that subject cannot see his own saccades in a mirror, that is there is in fact *saccadic suppression* (Matin, 1974).

The analysis on foveated-based neural computation nicely explains also the reason why humans cannot see video with a number of frames per second that exceeds the classic sampling

threshold. It turns out that this number is clearly connected with the velocity of the scan paths of the focus of attention. Of course, this is a computational issue which goes beyond biology and clearly affects machines as well.

**BOX 7** | The above items provide strong evidence for the reasons why foveate eyes turn out to be very effective for scene understanding. Interestingly, we can export the information-based principle of focussing attention to computer retinas by simulating eye movements. There is more: machines could provide multiple focuses of attention which could increase their visual skills significantly.

## 4.8 What Drives Eye Movements?

Foveate animals need to move their eyes to properly focus attention. The previous discussion has emphasized the importance of performing appropriate movements, which motivates the following question naturally arises:

**Q8:** *What are the mechanisms that drive eye movements?*

Human eyes make jerky saccadic movements during ordinary visual acquisition. One reason for these movements is that the fovea provides high-resolution in portions of about  $1, 2^\circ$ . Because of such a small high resolution portions, the overall sensing of a scene does require intensive movements of the fovea. Hence, the foveate movements do represent a good alternative to eyes with a uniformly high resolution retina. The information-based principles discussed so far lead us to conclude that foveate retinas with saccadic movements is in fact a solution that is computationally sustainable and very effective. Fast reactions to changes in the surrounding visual environment require efficient attention mechanisms to reallocate computational resources to most relevant locations in the visual field. While current computational models keep improving their predictive ability thanks to the increasing availability of data, they are still far away from the effectiveness and efficiency exhibited by foveate animals. An in-depth investigation on biologically-plausible computational models of focus of attention that exhibit spatiotemporal locality is very important also for computer vision, where one relies on parallel and distributed implementations. The research carried out by (Faggi et al., 2020) suggests an interpretation based on a computational model where attention emerges as a wave propagation process originated by visual stimuli corresponding to details and motion information. The resulting field obeys the principle of *inhibition of return*, so as not to get stuck in potential holes, and extend previous studies in (Zanca et al., 2020) with the main objective of providing spatiotemporal locality. In particular, the idea of modeling the focus of attention by a gravitational process finds its evolution in the corresponding local model based on the Poisson equation on the corresponding potential. Interestingly, Newtonian gravity yields an instantaneous propagation of signals, so as a sudden change in the mass density of a given pixel immediately affects the focus of attention, regardless of its location on the retina. These studies are driven by the principle that there are in fact sources which drive attention (e.g., masses in a gravitational field). At early cognitive

stages, attention mechanisms are mostly driven by the presence of details and movements. This is the reason why the mentioned masses for modeling the focus of attention have been based on the magnitude of the gradient of the brightness in the retina and on the optical flow. Interestingly, in children the mechanisms that drive the focus of attention are strongly connected with the developmental stages. Newborns and children in their early stages of evolution only focus attention on details and movements and on a few recurrent visual patterns like faces. As time goes by, visual features acquire a semantic value and, consequently, the focus of attention is gradually driven by specific intentions and corresponding plans. Of course, this is only possible after having acquired some preliminary capability of recognizing objects. Interestingly, as the forward process that facilitate high level cognitive tasks from the focus of attention becomes effective a corresponding backward process begins the improvement of the focus of attention. A reinforcement loop is generated which is finalized to optimize the final purpose of the agent in its own learning environment.

**BOX 8** | What drives the focus of attention is definitely a crucial issue, simply because of its already discussed fundamental role. We conjecture that this driving process must undergo a developmental process, where we begins with details and optical flow and proceed with the fundamental feedback from the environment which is clearly defined by the specific purpose of the agent.

## 4.9 Why is Baby Vision Blurred?

There are surprising results that come from developmental psychology on what a newborns see. Basically, their visual acuity grows gradually in early months of life. Interestingly, Charles Darwin had already noticed this very interesting phenomenon. In his words:

It was surprising how slowly he acquired the power of following with his eyes an object if swinging at all rapidly; for he could not do this well when seven and a half months old.

At the end of the seventies, this early remark was given a technically sound basis (see, e.g., Dobson and Teller 1978). In the paper, three techniques, — optokinetic nystagmus (OKN), preferential looking (PL), and the visually evoked potential (VEP)— were used to assess visual acuity in infants between birth and 6 months of age. More recently (Braddick and Atkinson, 2011), provides an in-depth discussion on the state of the art in the field. It is clearly stated that for newborns to gain adult visual acuity, depending on the specific visual test, several months are required. The following question naturally arises:

**Q9:** *Why does it take 8–12 months for newborns to achieve adult visual acuity? Is the development of adult visual acuity a biological issue or does it come from higher level computational laws of vision?*

This brings up the discussion on the “protection” of the learning agent from information overloading, which might be

of fundamental importance also in computer vision. The blurring of the video at an early stage of learning is compatible with a broader view of learning that, in addition to the involvement of the classic synaptic connections, this provides a direct “simplification of the input.” Regardless of this specific input modification, the underlying idea is that the process of learning consists of properly filtering the input with the purpose of gradually acquiring the information. The development of any computation model that adheres to this view is based on modifying the connections along with an appropriate input filtering so that the learning agent always operates at an equilibrium point (Betti et al., 2021).

**BOX 9 |** When promoting the role of time, the arising pre-algorithmic framework suggests extending the learning process to an appropriate modification of the input, that is finalized to achieve the expected “visual acuity” at the end of the process of learning. In doing so, one can think of generalization processes that are based on the convergence to fixed values of the weight connections, but also on an opportune “small perturbation” of the input.

#### 4.10 What is the Interplay With Language?

The interplay of vision and language is definitely one of the most challenging issues for an in-depth understanding of human vision. Along with the associated successes, the indisputable adoption of the supervised learning protocol in most challenging object recognition problems caused the losing of motivations for an in-depth understanding of the way linguistic information is synchronized with visual clues. In particular, the way humans learn the name of objects is far away from the current formal supervised protocol. This can likely be better grasped when we begin considering that top level visual skills can be found in many animals (e.g., birds and primates), which clearly indicates that their acquisition is independent of language.

Hence, as we clarify the interplay of vision and language we will likely address also the first question on how to overcome the need for “intensive artificial supervision.” Since first linguistic skills arise in children when their visual acuity is already very well developed, there is a good chance that early simple associations between objects and their names can easily be obtained by “a few supervisions” because of the very rich internal representation that has already been gained of those objects. It is in fact only a true independent hidden representation of objects which makes possible their subsequent association with a label! The capability of learning motion-invariance features is a fundamental information-based principle regardless of biology, which might somehow drives the development of “what” neurons.

The interplay of language and vision has been very well addressed in a survey by Lupyan (2012). It is claimed that performance on tasks that have been presumed to be non-verbal is rapidly modulated by language, thus rejecting the distinction between verbal and non-verbal representations. While we subscribe to the importance of sophisticated interactions, we also reinforce the claim that capturing the identity of single objects is mostly a visual issue. However, when we move towards the acquisition of abstract notions of

objects than the interaction with language is likely to be very important. One needs to separate single objects coming in visual contexts with their own identity with respect to abstract notions of objects. We can see a specific chair, but it’s a different story to recognize that we have a chair in front of us.

In **Section 4.7** we addressed the issue of motion invariance by claiming that it must properly be considered in the unified framework of focus of attention. We experience eye movements either on still images or during motion. In the first case we can see micro-saccadic movements, whereas moving objects are properly tracked. While the enforcement of visual feature invariance makes sense in both cases, there is a fundamental difference from an information-based viewpoint: The object tracking does provide information on the object movement, so that one can propagate the label to all the pixels that are connected with the pixel where we focus attention by a non-vanishing optical flow. This conveys an enormous amount of labelled information on the moving object and on its related segmentation. There is more! While during micro-saccadic movements many invariant features can be developed and it is not clear which one—if any—refers to an explicit object as a whole, during smooth pursuit, thanks to the optical flow, the moving object, with its own label, provides an internal representation gained under this motion invariance that is likely to be the secret for bridging the linguistic attachment of labels to objects.

The opportune exploitation of optical flow in visual information is of paramount importance for the evolution of theories of vision. In the last few years we have also seen a number of contributions in egocentric vision, where the assumption is that also the camera is moving. Interestingly, any sophisticated filtering of such an external movement might neglect the importance of undergoing developmental steps just like those that are fundamental for capturing the interplay with language. Clearly, if you have already gained good visual skills in object recognition, it is quite easy to check whether you yourself are moving!

Once again, the discussion carried out so far promotes the idea that for a visual agent to efficiently obtain the capabilities of recognizing objects from a few supervisions, it must undergo some developmental steps aimed at developing invariant representations of objects, so the actual linguistic supervision takes place only after the development of those representations. But, when should we enable a visual agent to begin with the linguistic interaction? While one might address this question when attacking the specific computational model under investigation, a more natural and interesting way to face this problem is to re-formulate the question as:<sup>4</sup>

**Q10:** *How can we develop “linguistic focusing mechanisms” that can drive the process of object recognition?*

<sup>4</sup>There’s not a morning I begin without a thousand questions running through my mind . . . The reason why a bird was given wings If not to fly, and praise the sky . . . –From Yentl, “Where is it Written?” –I.B. Singer, The Yeshiva Boy

This is done in a spectacular way in nature! Like vision, language development requires a lot of time. Interestingly, it looks like it requires more than vision. The discussion in **Section 4.9** indicates that the gradual growth of visual acuity is a possible clue to begin with language synchronization. The discussed filtering process offers a protection from visual information overloading that likely holds for language as well. As the visual acuity gradually increases, one immediately realizes that the mentioned visual-language synchronization has a spatiotemporal structure. At a certain time, we need to inform the agent about what we see at a certain position in the retina. An explicit implementation of such an association can be favored by an active learning process: the agent can ask itself what is located at  $(x, t)$ . However, what if you cannot rely on such a precious active interactions? For example, a linguistic description of the visual environment is generally very sophisticated and mentions objects located in different positions of the retina, without providing specific spatiotemporal information. Basically, this is a sort of *weak supervision* that is more difficult to grasp. However, once again, developmental learning schemes can significantly help. At early stage of learning the agent's tasks can be facilitated by providing spatiotemporal information. For example, naming the object located where the agent is currently focussing attention conveys information by a sort of human-like communication protocol. As time goes by, the agent gains gradually the capability of recognizing a few objects. What really matters is the confidence that is gained in such a task. When such a developmental stage is reached, linguistic descriptions and any sort of natural language based visual communication can be conveniently used to reinforce the agent recognition confidence. Basically, these weak supervisions turn out to be very useful since they can profitably be attached where the agent came up with a prediction that matches the supervision.

**BOX 10 |** Computer vision and natural language processing have been mostly evolving independently one each other. While this makes sense, the time has come to explore the interplay between vision and language with the main purpose of going beyond the protocol of supervised learning for attaching labels to objects. Interestingly challenges arises in scene interpretation when we begin considering the developmental stages of vision that suggest gaining strong object invariance before the attachment of linguistic labels.

## 5 THE “EN PLEIN AIR” PERSPECTIVE

Posing the right questions is the first fundamental step to gain knowledge and solve problems. The intent of this paper is to provide insights and to contribute to a shift in the direction in which computer vision is presently being practiced in the deep learning community. However, one might wonder what could be the most concrete action for promoting studies on the posed questions. So far, computer vision has strongly benefited from the massive diffusion of benchmarks which, by and large, are regarded as fundamental tools for performance evaluation. However, it is clear that they are very well-suited to support the statistical machine learning approach based on huge collections of labelled images. This paper, however, opens the doors to explore a different framework for performance

evaluation. The emphasis on video instead of images does not leads us to think of huge collection of video, but to adopt a different approach in which no collection at all is accumulated! Just like humans, machines are expected to live in their own visual environment. What should be the scientific framework for evaluating the performance and understand when a theory carries out important new results? Benchmarking bears some resemblance to the influential testing movement in psychology which has its roots in the turn-of-the-century work of Alfred Binet on IQ tests (Binet and Simon, 1916). Both cases consist of attempts to provide a rigorous way of assessing the performance or the aptitude of a (biological or artificial) system, by agreeing on a set of standardized tests which, from that moment onward, become the ultimate criterion for validity. On the other hand, it is clear that the skills of any visual agent can be quickly evaluated and promptly judged by humans, simply by observing its behavior. Thus, we could definitely rely on a *crowdsourcing performance evaluation scheme* where registered people can inspect and assess the performance of software agents (Gori et al., 2015). We use the *term en plein air* to mimic the French Impressionist painters of the 19th-century and, more generally, the act of painting outdoors. This term suggests that visual agents should be evaluated by allowing people to see them in action, virtually opening the doors of research labs. The *en plein air* proposal allows others to test our algorithms and to contribute to this evaluation method by providing their own data, their own results, or the comparisons with their own algorithms.

While the idea of shifting computer vision challenges into the wild will deserves attention one cannot neglect the difficulties that arise from the lack of a truly lab-like environment for supporting the experiments. The impressive progress in computer graphics, however, offers a very attractive alternative that can dramatically facilitate the developments of approaches to computer vision that are based on the on-line treatment of the video (see, e.g., Meloni et al., 2020).

Needless to say, computer vision has been fueled by the availability of huge labelled image collections, which clearly shows the fundamental role played by pioneer projects in this direction (see, e.g., Deng et al., 2009). The ten questions posed in this paper will likely be better addressed only when scientists will put more emphasis on the *en plein air* environment. In the meantime, the major claim of this paper is that the experimental setting needs to move to virtual visual environments. Their photorealistic level along with the explosion of the generative capabilities makes these environments better suited to new performance evaluation of computer vision.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Files, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.



## REFERENCES

- Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active Vision. *Int. J. Comput. Vis.* 1, 333–356. doi:10.1007/bf00133571
- Baker, S., Scharstein, D., Lewis, J. P., Roth, S., Black, M. J., and Szeliski, R. (2011). A Database and Evaluation Methodology for Optical Flow. *Int. J. Comput. Vis.* 92, 1–31. doi:10.1007/s11263-010-0390-2
- Baker, N., Erlikhman, G., Kellman, P. J., and Lu, H. (2018). “Deep Convolutional Networks Do Not Perceive Illusory Contours,” in Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25–28, 2018. Editors C. Kalish, M. A. Rau, X. J. Zhu, and T. T. Rogers. cognitivesciencesociety.org.
- Ballard, D. H. (1991). Animate Vision. *Artif. Intell.* 48, 57–86. doi:10.1016/0004-3702(91)90080-4
- Beltran, W. A., Cideciyan, A. V., Guziewicz, K. E., Iwabe, S., Swider, M., Scott, E. M., et al. (2014). Canine Retina Has a Primate Fovea-like Bouquet of Cone Photoreceptors Which Is Affected by Inherited Macular Degenerations. *PLOS ONE* 9, 1–10. doi:10.1371/journal.pone.0090390
- Benjamin, W. T., Mary, M. H., Michael, F. L., and Dana, H. B. (2011). Eye Guidance in Natural Vision: Reinterpreting Saliency. *J. Vis.* 11, 1–23. doi:10.1167/11.5.5
- Bertasius, G., Wang, H., and Torresani, L. (2021). *Is Space-Time Attention All You Need for Video Understanding?* arXiv:2102.05095
- Betti, A., and Gori, M. (2016). The Principle of Least Cognitive Action. *Theor. Comput. Sci.* 633, 83–99. doi:10.1016/j.tcs.2015.06.042
- Betti, A., and Gori, M. (2018). *Convolutional Networks in Visual Environments*. Arxiv preprint arXiv:1801.07110v1.
- Betti, A., and Gori, M. (2019). *Backprop Diffusion Is Biologically Plausible*. CoRR abs/1912.04635.
- Betti, A., Gori, M., and Melacci, S. (2018). *Cognitive Action Laws: The Case of Visual Features*. CoRR abs/1808.09162
- Betti, A., Gori, M., and Melacci, S. (2021). *Learning and Visual Blurring*. Technical Report. SAILab.
- Binet, A., and Simon, T. (1916). *The Development of Intelligence in Children: The Binet-Simon Scale*. Williams & Wilkins.
- Borenstein, E., and Ullman, S. (2002). “Class-specific, top-down segmentation,” in *Computer Vision - ECCV 2002, 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28–31, 2002, Proceedings, Part II. Lecture Notes in Computer Science*. Editors A. Heyden, G. Sparr, M. Nielsen, and P. Johansen (Springer), 2351, 109–124. doi:10.1007/3-540-47967-8\_8
- Braddick, O., and Atkinson, J. (2011). Development of Human Visual Function. *Vis. Res.* 51, 1588–1609. doi:10.1016/j.visres.2011.02.018
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). “ImageNet: A Large-Scale Hierarchical Image Database,” in CVPR09, Miami, FL, June 22–24, 2009. doi:10.1109/cvpr.2009.5206848
- Dobson, V., and Teller, D. Y. (1978). Visual Acuity in Human Infants: A Review and Comparison of Behavioral and Electrophysiological Studies. *Vis. Res.* 18, 1469–1483. doi:10.1016/0042-6989(78)90001-9
- Faggi, L., Betti, A., Zanca, D., Melacci, S., and Gori, M. (2020). *Wave Propagation of Visual Stimuli in Focus of Attention*. CoRR abs/2006.11035.
- Gerkema, M., Davies, W., Foster, R., Menaker, M., and Hut, R. (2013). The Nocturnal Bottleneck and the Evolution of Activity Patterns in Mammals. *Proc. R. Soc. Lond. Ser. B, Biol. Sci.* 280, 20130508. doi:10.1098/rspb.2013.0508
- Gibson, J. J., and Boston (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
- Gibson, J. J. (1950). *The Perception of the Visual World*. Houghton Mifflin.
- Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.
- Goodale, M. A., and Keith Humphrey, G. (1998). The Objects of Action and Perception. *Cognition* 67, 181–207. doi:10.1016/s0010-0277(98)00017-1
- Goodale, M. A., and Milner, A. D. (1992). Separate Visual Pathways for Perception and Action. *Trends Neurosci.* 15, 20–25. doi:10.1016/0166-2236(92)90344-8
- Gori, M., Lippi, M., Maggini, M., Melacci, S., and Pelillo, M. (2015). “En plein air visual agents,” in *Image Analysis and Processing - ICIAP 2015 - 18th International Conference, Genoa, Italy, September 7–11, 2015, Proceedings, Part II. Lecture Notes in Computer Science*. Editors V. Murino and E. Puppo (Springer), 9280, 697–709. doi:10.1007/978-3-319-23234-8\_64
- Gori, M., Lippi, M., Maggini, M., and Melacci, S. (2016). Semantic Video Labeling by Developmental Visual Agents. *Computer Vis. Image Understanding* 146, 9–26. doi:10.1016/j.cviu.2016.02.011
- Goroshin, R., Bruna, J., Tompson, J., Eigen, D., and LeCun, Y. (2015). “Unsupervised Learning of Spatiotemporally Coherent Metrics,” in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015, 4086–4093. doi:10.1109/ICCV.2015.465
- Horn, B. K., and Schunck, B. (1981). Determining Optical Flow. *Artif. Intell.* 17, 185–203. doi:10.1016/0004-3702(81)90024-2
- Hubel, D., and Wiesel, T. (1962). Receptive fields, Binocular Interaction, and Functional Architecture in the Cat’s Visual Cortex. *J. Physiol. (London)* 160, 106–154. doi:10.1113/jphysiol.1962.sp006837
- Kim, B., Reif, E., Wattenberg, M., and Bengio, S. (2019). *Do Neural Networks Show Gestalt Phenomena? an Exploration of the Law of Closure*. CoRR abs/1903.01069.
- Kingstone, A., Daniel, S., and John, D. E. (2010). Cognitive Ethology: A New Approach for Studying Human Cognition. *Br. J. Psychol.* 99, 317–340. doi:10.1348/000712607x251243
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature* 521, 436–444. doi:10.1038/nature14539
- Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). “Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations,” in Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09 (New York, NY, USA: ACM), 609–616. doi:10.1145/1553374.1553453
- Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. (1959). What the Frog’s Eye Tells the Frog’s Brain. *Proc. IRE* 47, 1940–1951. doi:10.1109/jrproc.1959.287207
- Lupyan, G. (2012). Linguistically Modulated Perception and Cognition: The Label-Feedback Hypothesis. *Front. Psychol.* 3, 54. doi:10.3389/fpsyg.2012.00054
- Marinai, S., Gori, M., and Soda, G. (2005). Artificial Neural Networks for Document Analysis and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 23–35. doi:10.1109/tpami.2005.4
- Marr, D. (1982). *Vision*. San Francisco: Freeman. Partially reprinted in Anderson and Rosenfeld (1988).
- Matin, E. (1974). Saccadic Suppression: A Review and an Analysis. *Psychol. Bull.* 81, 899–917. doi:10.1037/h0037368
- Meloni, E., Pasqualini, L., Tiezzi, M., Gori, M., and Melacci, S. (2020). *Sailenv: Learning in Virtual Visual Environments Made Simple*. CoRR abs/2007.08224.
- Ott, J., Linstead, E., LaHaye, N., and Baldi, P. (2020). Learning in the Machine: To Share or Not to Share? *Neural Networks* 126, 235–249. doi:10.1016/j.neunet.2020.03.016
- Pan, S., and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi:10.1109/tkde.2009.191
- Poggio, T. A., and Anselmi, F. (2016). *Visual Cortex and Deep Networks: Learning Invariant Representations*. 1st edn. The MIT Press.
- Ranzato, M., Huang, F. J., Boureau, Y., and LeCun, Y. (2007). “Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition,” in 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, Minnesota, USA, 18–23 June 2007. doi:10.1109/CVPR.2007.383157
- Rao, R. P. N., and Ballard, D. H. (1999). Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nat. Neurosci.* 2, 79–87. doi:10.1038/4580
- Ronneberger, O., Fischer, P., and Brox, T. (2015). *U-net: Convolutional Networks for Biomedical Image Segmentation*. CoRR abs/1505.04597. doi:10.1007/978-3-319-24574-4\_28
- Ross, C. F. (2004). *The Tarsier Fovea: Functionless Vestige or Nocturnal Adaptation?* Boston, MA: Springer US, 477–537. doi:10.1007/978-1-4419-8873-7\_19
- Sohn, E. (2019). The Eyes of Mammals Reveal a Dark Past. *Nature*. doi:10.1038/d41586-019-01109-6
- Tavanaei, A., Masquelier, T., and Maida, A. S. (2016). *Acquisition of Visual Features through Probabilistic Spike-timing-dependent Plasticity*. CoRR abs/1606.01102. doi:10.1109/ijcnn.2016.7727213
- Ullman, S. (1979). *The Interpretation of Visual Motion/Shimon Ullman*. The MIT press series in artificial intelligence (The MIT press).
- Walls, G. L. (1942). *The Vertebrate Eye and its Adaptive Radiation*.
- Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical*. USA: John Wiley & Sons.
- Wood, J. N., and Wood, S. M. (2020). One-shot Learning of View-Invariant Object Representations in Newborn Chicks. *Cognition* 199, 104192. doi:10.1016/j.cognition.2020.104192

- Wood, J. N. (2016). A Smoothness Constraint on the Development of Object Recognition. *Cognition* 153, 140–145. doi:10.1016/j.cognition.2016.04.013
- Zanca, D., Melacci, S., and Gori, M. (2020). Gravitational Laws of Focus of Attention. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2983–2995. doi:10.1109/TPAMI.2019.2920636
- Zitnick, C. L., and Dollár, P. (2014). “Edge Boxes: Locating Object Proposals from Edges,” in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, 391–405. doi:10.1007/978-3-319-10602-1\_26

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gori. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Outer Product-Based Fusion of Smartwatch Sensor Data for Human Activity Recognition

Adria Mallol-Ragolta<sup>1\*</sup>, Anastasia Semertzidou<sup>1</sup>, Maria Pateraki<sup>2,3</sup> and Björn Schuller<sup>1,4</sup>

<sup>1</sup> EHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany,

<sup>2</sup> Institute of Computer Science, Foundation of Research and Technology – Hellas, Heraklion, Greece, <sup>3</sup> School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Athens, Greece, <sup>4</sup> GLAM – Group on Language, Audio, & Music, Imperial College London, London, United Kingdom

## OPEN ACCESS

### Edited by:

Pekka Siirtola,  
University of Oulu, Finland

### Reviewed by:

Saeed Hamood Alsamhi,  
Ibb University, Yemen  
Martin Gjoreski,  
University of Italian Switzerland,  
Switzerland

### \*Correspondence:

Adria Mallol-Ragolta  
adria.mallol-ragolta@  
informatik.uni-augsburg.de

### Specialty section:

This article was submitted to  
Mobile and Ubiquitous Computing,  
a section of the journal  
Frontiers in Computer Science

**Received:** 17 November 2021

**Accepted:** 18 February 2022

**Published:** 22 March 2022

### Citation:

Mallol-Ragolta A, Semertzidou A,  
Pateraki M and Schuller B (2022)  
Outer Product-Based Fusion of  
Smartwatch Sensor Data for Human  
Activity Recognition.  
Front. Comput. Sci. 4:796866.  
doi: 10.3389/fcomp.2022.796866

The advent of IoT devices in combination with Human Activity Recognition (HAR) technologies can contribute to battle with sedentariness by continuously monitoring the users' daily activities. With this information, autonomous systems could detect users' physical weaknesses and plan personalized training routines to improve them. This work investigates the multimodal fusion of smartwatch sensor data for HAR. Specifically, we exploit pedometer, heart rate, and accelerometer information to train unimodal and multimodal models for the task at hand. The models are trained end-to-end, and we compare the performance of dedicated Recurrent Neural Network-based (RNN) and Convolutional Neural Network-based (CNN) architectures to extract deep learnt representations from the input modalities. To fuse the embedded representations when training the multimodal models, we investigate a concatenation-based and an outer product-based approach. This work explores the harAGE dataset, a new dataset for HAR collected using a Garmin Vivoactive 3 device with more than 17 h of data. Our best models obtain an Unweighted Average Recall (UAR) of 95.6, 69.5, and 60.8 % when tackling the task as a 2-class, 7-class, and 10-class classification problem, respectively. These performances are obtained using multimodal models that fuse the embedded representations extracted with dedicated CNN-based architectures from the pedometer, heart rate, and accelerometer modalities. The concatenation-based fusion scores the highest UAR in the 2-class classification problem, while the outer product-based fusion obtains the best performances in the 7-class and the 10-class classification problems.

**Keywords:** artificial intelligence, human activity recognition, multimodal fusion, ubiquitous computing, smartwatch sensor data

## 1. INTRODUCTION

According to the *World Health Organization* (WHO), physical inactivity is a serious public health concern with serious implications in people's health, as it can be a risk factor for diabetes, depression, high blood pressure, or obesity. Physical activity is beneficial not only for physical health, but also for wellbeing (Fox, 1999; Penedo and Dahn, 2005). Hence, there is a need to develop new, digital, and personalized tools that engage their users to exercise with the goal to have a more active, and healthier life. The research performed on the field of *Human Activity Recognition* (HAR) can contribute to achieve this goal. This field of knowledge aims to develop technologies able to

recognize and, therefore, monitor the activities that users do. The exploitation of this information has a wide range of applications in many different domains, such as healthcare, fitness, athletics, elderly care, security, or entertainment.

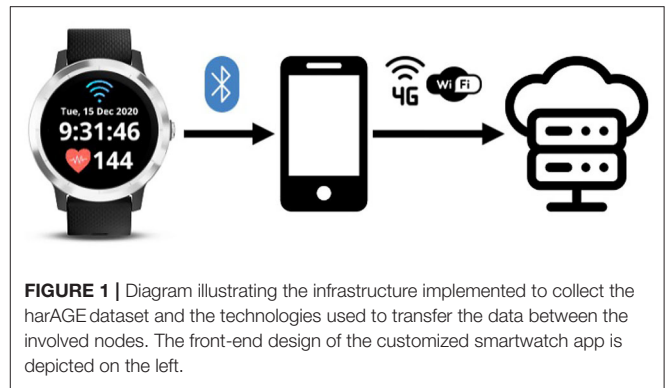
Commercial smartphones are equipped with embedded sensors, including accelerometers and gyroscopes, which make them suitable to recognize human activities (Khan et al., 2010; Bayat et al., 2014; Chen and Shen, 2017). Previous works in the literature explored different machine learning techniques, such as hidden Markov models (Ronao and Cho, 2014), unsupervised learning (Kwon et al., 2014), and deep learning (Ronao and Cho, 2016; Hassan et al., 2018) for HAR using smartphone sensor data. Smartwatches are a high-potential device for this task as well (Weiss et al., 2016; Shahmohammadi et al., 2017; Mekruksavanich and Jitpattanakul, 2020) because of their market penetration in society, which is increasing every year, and the embedded sensors they contain, which can be used to retrieve pedometer, photoplethysmographic, and accelerometer measurements (Lara et al., 2012). Furthermore, their location on the users' wrist seems advantageous to capture human activities.

This work investigates the multimodal fusion of pedometer, heart rate, and accelerometer information to train end-to-end models for HAR. One of the goals of this work is to determine which modalities are more suitable to be fused for the task at hand. Based on the tensor fusion layer presented by Zadeh et al. (2017), we propose using an outer product-based approach to fuse the embedded representations of the input modalities. The performance of this approach is compared with a concatenation-based approach, which we use as a baseline. The embedded representations of the input modalities are learnt using dedicated *Recurrent Neural Network*-based (RNN) or *Convolutional Neural Network*-based (CNN) architectures. Our experiments explore the harAGE dataset, a new smartwatch-based HAR dataset collected using a Garmin Vivoactive 3 device. The dataset contains more than 17 h of data from 19 participants while lying, sitting, standing, washing hands, walking, running, climbing stairs, doing strength and flexibility workout activities, and cycling.

The rest of this article is laid out as follows. Section 2 first highlights relevant related works in the literature. Section 3 presents the dataset employed, while Section 4 describes the methodology followed. Section 5 analyzes the results obtained from the experiments performed, and Section 6 concludes this article and suggests some future work directions.

## 2. RELATED WORK

The problem of HAR is an active topic in the research community. A large body of knowledge has tackled the problem from a computer vision perspective (Khaire et al., 2018; Qi et al., 2018), exploiting color, depth, and even skeletal information. We can consider these as passive approaches, as they require cameras overseeing the scene to perform inferences. On the other side, we can consider as active approaches those that use body-worn sensors for recognizing human activities. In this case, the sensors themselves experience the activities, and, therefore, the



sensor measurements can be directly used to infer them. Research on this topic has been conducted using dedicated heart rate sensors (Tapia et al., 2007), inertial/magnetic sensors (Altun and Barshan, 2010), or accelerometer sensors (Lin et al., 2018).

A wide range of sensors are embedded in consumer, smart devices nowadays, including smartphones and smartwatches. Their high penetration in society has motivated the use of data collected with such devices for HAR purposes (Ahmed et al., 2020; Ashry et al., 2020; Mekruksavanich and Jitpattanakul, 2020; Wan et al., 2020). From a user-centered perspective, the field of HAR has traditionally focused on recognizing the activities individuals do. Nevertheless, recent works are considering the problem from a *Multi-user Activity Recognition* (MAR) perspective, which addresses the activities that a group of individuals do to achieve a common goal (Li et al., 2020).

Multimodal approaches have been used in a wide variety of problems and applications to complement and enrich the information embedded in a single modality. Different fusion techniques, from simple to complex, have been explored for this purpose. Examples of simple fusion techniques include the element-wise sum or product of the features extracted from different modalities, or even their simple concatenation. Among the more complex techniques, researchers have investigated circulant fusion (Wu and Han, 2018), gated fusion (Kim et al., 2018), memory (Priyasad et al., 2021), graph neural networks (Holzinger et al., 2021), and even transformers (Prakash et al., 2021).

## 3. DATASET

This work explores the first version of harAGE: a new smartwatch-based dataset for HAR collected using a customized smartwatch app running on a Garmin Vivoactive 3 device (Mallol-Ragolta et al., 2021). The app reads the accelerometer, the heart rate, and the pedometer information available from the built-in embedded sensors. While the accelerometer information is sensed at 25 Hz, the sampling rate of the heart rate and the steps information is 1 Hz. The back-end of the smartwatch app encapsulates the data into a JSON message, which is sent in close to real-time into a customized, encrypted, and secure server via the Internet using the HTTPS protocol (cf. Figure 1).



**TABLE 1** | Summary of the activities included in the harAGE dataset, the number of participants collected for each activity, and the amount of data available time-wise.

Activity	Participants	Duration (HH):MM:SS
Resting	19	1:25:24
Lying	19	1:39:01
Sitting	18	1:31:25
Standing	18	1:35:51
Washing hands	18	53:40
Walking	18	2:23:59
Running	16	1:58:28
Stairs climbing	18	2:17:23
Strength Workout	18	53:05
Flexibility Workout	18	56:50
Cycling	13	1:36:40
$\Sigma$	19	17:11:46

The recruited participants followed a protocol especially designed for the collection of the harAGE dataset. The participants started with a resting phase during 5 min to collect their heart rate at rest, avoiding stressors and external stimuli. This measurement can be used as the baseline heart rate for each individual participant. Then, they performed a sequence of static activities including lying, sitting, and standing. These three activities were performed twice: first without moving, and then allowing reasonable free movements. Each one of these activities was performed during 3 min. Next, we asked participants to simulate washing their hands, without running water, also for 3 min. Although this activity was rarely included in previous HAR datasets found in the literature, the current pandemic context and the favorable placement of the smartwatch in the participants' wrist motivated its inclusion in the data collection protocol.

The following dynamic activities were included next in the protocol: walking, running, climbing stairs (both upstairs and downstairs), and cycling. Furthermore, each one of these activities was performed three times at low, moderate, and high intensities during 3 min each. Intensity levels are subjective, as these depend on several factors, such as the previous physical condition of the participants. Thus, to capture this variability in our dataset, we relied on the participants themselves to set their own thresholds for each intensity level. Before the cycling set of activities, we incorporated a set of workout activities in the protocol. These activities included two sets of strength workout activities (squats and arm raising exercises), and two sets of flexibility workout activities (shoulder roll and wrist stretching exercises). These four activities were performed for 1.5 min each.

To guarantee the safety measures against the COVID-19 pandemic, the dataset was mainly collected outdoors. This scenario posed a challenge, as the data transfer between the smartwatch and the server when the participants were outdoors was performed via the 4G connection of the smartphone

with which the smartwatch was paired. The back-end of the smartwatch app discards the old measurements unsuccessfully sent to the server as a preventive measure to avoid running out of memory because of an overflow of the internal buffers implemented to temporarily store the sensed measurements before being transmitted. As the 4G connection might slow down the data transmission, the amount of measurements buffered might be larger and, therefore, prone to losses. This was the reason why the measurements received from each activity occasionally contained discontinuities. As a pre-processing stage and to ensure the continuous stream of information, we trimmed the received data into segments of at least 20 s of consecutive sensor measurements. These segments are then used to populate the dataset.

This first version of the harAGE dataset contains 17 h 11 min 46 s of data from 19 participants (9 f, 10 m), with a mean age of 41.73 years and a standard deviation of 7.97 years. Before the data collection, participants read and signed an *Informed Consent Form* (ICF), which was previously approved by the competent ethics committee. A summary of the different activities considered in the dataset, and the amount of data available for each activity is provided in **Table 1**. Some participants partially completed the activities included in the protocol because of data transmission issues, or the impossibility to get access to a bike for the cycling-related activities.

## 4. METHODOLOGY

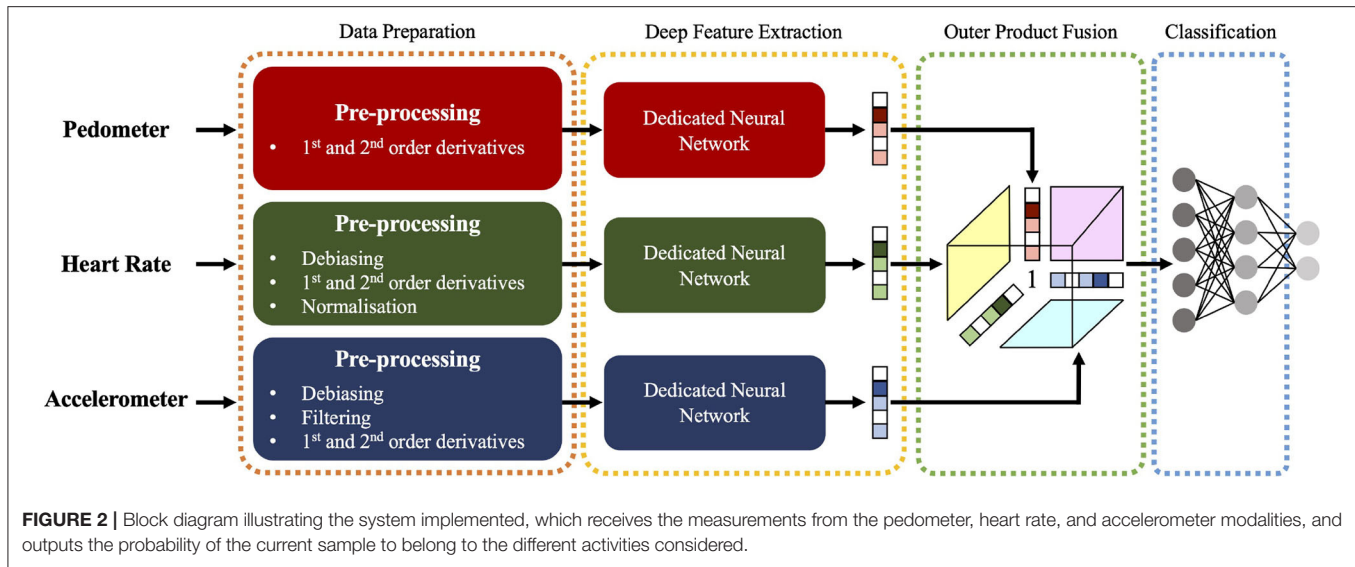
This section presents the methodology followed in this work to train end-to-end models for HAR using multimodal smartwatch data (cf. **Figure 2**). Section 4.1 describes the pre-processing applied to the raw measurements, Section 4.2 introduces the models implemented, and Section 4.3 summarizes their training details.

### 4.1. Data Preparation

In this passage, we describe the pre-processing applied to the raw measurements, which is different for each modality. After the pre-processing stage, the resulting information from each modality is segmented using windows of 20 s length and a 50 % overlap or without overlap, depending on whether the data is used for training or testing purposes, respectively. While each window contains 20 data points for the pedometer and the heart rate measurements, it contains 500 data points for the accelerometer measurements. The pre-processing applied to each modality is described below.

#### 4.1.1. Pedometer Measurements

The Garmin Vivoactive 3 device allows retrieving the number of steps performed by the user since midnight. The absolute number of steps is not a suitable feature to model the current human activity, as the cumulative effect caused by the nature of the embedded sensor conditions the measurements. For instance, a high number of steps does not necessarily mean that the user is currently exercising, as the physical activity might have taken place a while—or even a long—ago. Instead, we hypothesize that the first and the second order derivatives computed from



the absolute number of steps could be more suitable features to characterize the users' activities, as these could model the velocity and the acceleration of the users' steps instantaneously. Hence, in our experiments, we use the first and the second order derivatives of the pedometer information as the features to extract from this modality.

#### 4.1.2. Heart Rate Measurements

The characteristics of the human heart while exercising are person-dependent, as they might depend on a wide range of variables, including age, physical condition, or existing pathologies, among others. To remove this personal bias from our data, we compute the median of the heart rates collected from each participant during the resting activity individually and use this measurement as the personal, baseline heart rate. We opt for computing the median to avoid considering the outliers in the raw measurements. The heart rate measurements collected from all the activities performed by each participant are debiased using the corresponding personal, baseline heart rate. For this modality, we also compute the first and the second order derivatives of the debiased heart rate signals in order to better characterize their dynamics over time. Finally, we normalize the debiased heart rate signal by a factor of 220 BPM (beats per minute), which is widely considered as the maximum heart rate of a human being. Although the maximum heart rate is age-dependent from a theoretical point of view (Fox and Naughton, 1972), we disregard this factor and apply the same normalization parameter to all participants in the dataset. Therefore, in our experiments, we use the first and the second order derivatives of the debiased heart rate signals, and their normalized representation as the features to extract from this modality.

#### 4.1.3. Accelerometer Measurements

It is sometimes possible to identify a person just by the way how she or he walks or moves. This observation leads us

to hypothesize that the accelerometer measurements collected using a smartwatch can contain personal information that might interfere in the intrinsic movements of the activities considered in the harAGE dataset. To overcome this issue, we first read all the accelerometer measurements available in the dataset for each individual user separately and compute the median of the measurements in the  $x$ -,  $y$ -, and  $z$ -axes. We then use this information to debias the raw accelerometer measurements in a personalized manner. A 1-dimensional Gaussian filter, using a Gaussian kernel with a standard deviation of 1, is used to remove noises and smooth the accelerometer measurements in the 3 different axes separately (Zhuang and Xue, 2019). We finally compute the first and the second order derivatives of the debiased, filtered accelerometer measurements in order to better characterize the dynamics of this modality over time. Thus, in our experiments, we use the debiased, filtered accelerometer measurements and their first and second order derivatives as the features to extract from each axis of this modality.

### 4.2. Models Descriptions

The end-to-end models implemented in this work are composed of three different blocks: (i) the first block extracts dedicated deep learnt representations from the modality-dependent sequence of features defined in Section 4.1 (cf. Section 4.2.1), (ii) the second block, which is enabled when training multimodal models only, is in charge of fusing the embedded representations of the modalities selected (cf. Section 4.2.2), and (iii) the third and final block is responsible for performing the actual classification (cf. Section 4.2.3).

#### 4.2.1. Deep Features Extraction

This block in the architecture is modality-specific; i.e., a dedicated feature extraction block processes the sequences of features from each modality separately. We compare two different network architectures for this task: an RNN, and a CNN. We use RNNs and CNNs as deep feature extractors,

**TABLE 2 |** Summary of the descriptive statistics ( $\mu$ : mean,  $\sigma$ : standard deviation) computed from the UAR scores obtained when assessing the unimodal and the multimodal binary classification-based end-to-end models using nested LOSO-CV.

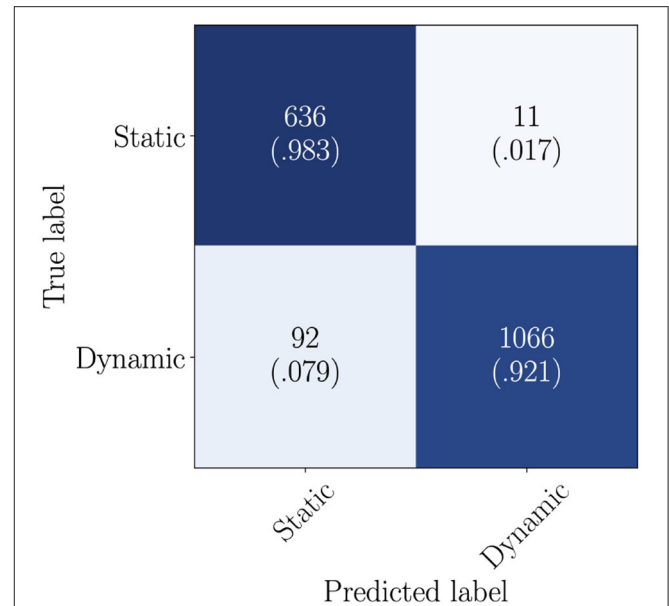
UAR [%]	RNN		CNN	
	$\mu$	$\sigma$	$\mu$	$\sigma$
$f_{steps}$	88.6	9.7	88.4	9.2
$f_{hr}$	82.8	15.2	85.2	12.4
$f_{xyz}$	70.2	19.3	90.9	13.5
$f_{steps@hr}$	91.0	13.0	94.7	8.4
$f_{steps@hr}$	92.9	7.2	91.8	13.7
$f_{steps@xyz}$	89.3	13.2	91.5	13.2
$f_{steps@xyz}$	88.6	12.8	88.3	11.4
$f_{hr@xyz}$	85.8	14.4	95.2	8.3
$f_{hr@xyz}$	86.6	14.1	92.6	13.2
$f_{steps@hr@xyz}$	<b>93.7</b>	6.4	<b>95.6</b>	5.5
$f_{steps@hr@xyz}$	90.5	15.1	95.4	4.6

The results compare the use of an RNN-based and a CNN-based architecture to extract deep learnt representations from the input modalities, and the fusion of the embedded representations in the multimodal models using a concatenation-based (represented with  $\oplus$ ) and an outer product-based (represented with  $\otimes$ ) approach. The bold values highlight the best results using each architecture.

since they have been extensively used in the literature for such purpose. The RNN implements a single layer, bidirectional *Gated Recurrent Unit-Recurrent Neural Network* (GRU-RNN) with 8 hidden units. The CNN implements a single 1-dimensional convolutional layer with 8 filters, a kernel size of 2, and a stride of 1. Following this convolutional layer, we use 1-dimensional batch normalization, and the output is transformed using a *Rectified Linear Unit* (ReLU) function. A 1-dimensional adaptive average pooling layer is implemented at the end of this convolutional block, so it produces 2 features per filter. The parameters of the RNN- and the CNN-based architectures are designed, so they both produce 16 deep learnt features at the output. This way, we can fairly compare the performances between both approaches. The dimensionality of the deep learnt features is also engineered, so the resulting embeddings from the outer product-based fusion when training the multimodal models have a reasonable dimensionality in terms of computational cost.

#### 4.2.2. Multimodal Fusion

One of the goals of this work is to investigate the suitability of using an outer product-based approach to fuse the embedded presentations learnt from different modalities in the problem of HAR. As a baseline, we use the simplest fusion method: the inner concatenation of the deep learnt representations from each modality. Representing these embedded representations for the pedometer, heart rate, and accelerometer modalities as  $f_{steps}$ ,  $f_{hr}$ , and  $f_{xyz}$ , respectively, we mathematically define the concatenation-based fusion as:

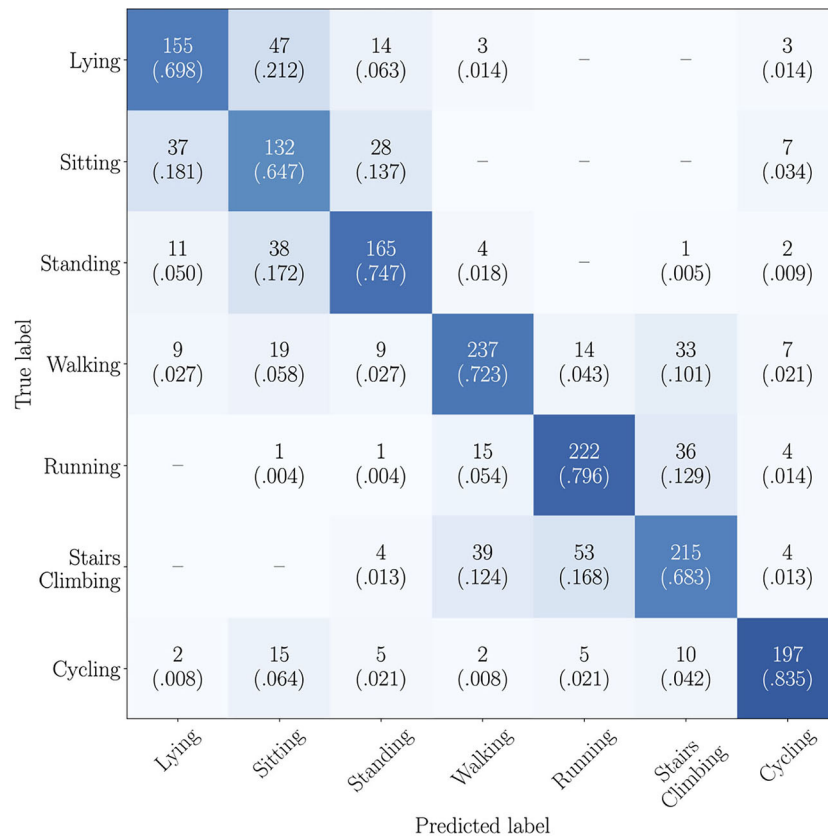


**FIGURE 3 |** Confusion matrix computed from the predictions inferred using the RNN-based approach exploiting  $f_{steps@hr@xyz}$  as input features. The static class includes the lying, sitting, and standing activities, while the dynamic class includes the walking, running, stairs climbing, and cycling activities. Each cell contains the absolute and the relative number of samples of the actual class (row) classified into each possible activity of the recognition set (column). The relative information is also depicted using a color-scale: dark colors illustrate high percentages, while light colors, low percentages.

**TABLE 3 |** Summary of the descriptive statistics ( $\mu$ : mean,  $\sigma$ : standard deviation) computed from the UAR scores obtained when assessing the unimodal and the multimodal standard HAR-based end-to-end models using nested LOSO-CV.

UAR [%]	RNN		CNN	
	$\mu$	$\sigma$	$\mu$	$\sigma$
$f_{steps}$	30.2	11.1	30.7	5.6
$f_{hr}$	32.9	5.3	34.6	5.9
$f_{xyz}$	40.6	16.5	56.9	14.9
$f_{steps@hr}$	51.4	10.1	47.7	13.7
$f_{steps@hr}$	55.8	12.5	50.7	12.8
$f_{steps@xyz}$	53.7	16.1	52.3	10.7
$f_{steps@xyz}$	58.2	11.4	58.2	7.3
$f_{hr@xyz}$	59.9	15.6	62.8	16.8
$f_{hr@xyz}$	<b>62.5</b>	12.6	68.3	13.7
$f_{steps@hr@xyz}$	59.7	14.8	<b>69.5</b>	14.4
$f_{steps@hr@xyz}$	57.0	11.3	<b>69.5</b>	12.2

The results compare the use of an RNN-based and a CNN-based architecture to extract deep learnt representations from the input modalities, and the fusion of the embedded representations in the multimodal models using a concatenation-based (represented with  $\oplus$ ) and an outer product-based (represented with  $\otimes$ ) approach. The bold values highlight the best results using each architecture.



**FIGURE 4 |** Confusion matrix computed from the predictions inferred using the CNN-based approach exploiting  $f_{steps \otimes hr \otimes xyz}$  as input features. Each cell contains the absolute and the relative number of samples of the actual class (row) classified into each possible activity of the recognition set (column). The relative information is also depicted using a color-scale: dark colors illustrate high percentages, while light colors, low percentages. Empty cells indicate no samples from the actual activity are classified into the corresponding class.

$$f_{steps \oplus hr \oplus xyz} = \begin{bmatrix} f_{steps} \\ f_{hr} \\ f_{xyz} \end{bmatrix}. \quad (1)$$

The dimensionality of the resulting embedded representation from the concatenation-based fusion is  $\mathbb{R}^{16 \times m}$ , where  $m$  indicates the number of modalities to be fused. When all three modalities are fused together ( $m = 3$ ), the resulting embedded representation is  $\in \mathbb{R}^{48}$ . The outer product-based fusion proposed is inspired by the tensor fusion layer presented by Zadeh et al. (2017) and can be mathematically defined as:

$$f_{steps \otimes hr \otimes xyz} = \begin{bmatrix} f_{steps} \\ 1 \end{bmatrix} \otimes \begin{bmatrix} f_{hr} \\ 1 \end{bmatrix} \otimes \begin{bmatrix} f_{xyz} \\ 1 \end{bmatrix}. \quad (2)$$

When the three modalities are fused together, the outer product generates a cube with the following properties: (i) the original representations are preserved in the edges of the cube, (ii) each face of the cube contains information from the fusion of two modalities, and (iii) the inner part of the cube fuses information from the three modalities all together. The fused representation

is flattened before being fed into the final, classification block of the network. The dimensionality of the resulting embedded representation from the outer product-based fusion is  $\mathbb{R}^{(16+1)^m}$ . When all three modalities are fused together ( $m = 3$ ), the resulting embedded representation is  $\in \mathbb{R}^{4913}$ .

#### 4.2.3. Classification

The classification block of the network implements two fully connected layers, preceded by a dropout layer with probability 0.3. The number of input neurons in the first fully connected layer depends on the number of modalities to be fused during the training process. The output of this layer produces a 16-dimensional representation, which is transformed using a ReLU activation function. This transformed representation is fed into the second fully connected layer, which contains as many neurons at the output as activities we need to classify our samples into, and uses a Softmax activation function. This way, the network outputs can be interpreted as probability scores.

#### 4.3. Networks Training

For a fair comparison among the models, these are all trained under the exact same conditions. The pseudorandom number



**TABLE 4 |** Summary of the descriptive statistics ( $\mu$ : mean,  $\sigma$ : standard deviation) computed from the UAR scores obtained when assessing the unimodal and the multimodal multi-class harAGE-based end-to-end models using nested LOSO-CV.

UAR [%]	RNN		CNN	
	$\mu$	$\sigma$	$\mu$	$\sigma$
$f_{steps}$	21.1	7.7	20.7	8.3
$f_{hr}$	23.6	3.9	24.1	4.7
$f_{xyz}$	34.3	11.7	43.4	12.5
$f_{steps@hr}$	38.2	9.1	33.2	9.1
$f_{steps@hr}$	38.5	9.2	35.8	9.1
$f_{steps@xyz}$	46.5	10.4	41.3	11.1
$f_{steps@xyz}$	47.4	10.5	51.0	12.5
$f_{hr@xyz}$	46.3	12.9	46.8	11.6
$f_{hr@xyz}$	48.7	12.9	60.5	13.6
$f_{steps@hr@xyz}$	<b>56.6</b>	10.1	54.8	13.3
$f_{steps@hr@xyz}$	49.7	9.6	<b>60.8</b>	12.8

The results compare the use of an RNN-based and a CNN-based architecture to extract deep learnt representations from the input modalities, and the fusion of the embedded representations in the multimodal models using a concatenation-based (represented with  $\oplus$ ) and an outer product-based (represented with  $\otimes$ ) approach. The bold values highlight the best results using each architecture.

generator is seeded at the initialization of the models for reproducibility purposes. The networks are trained to minimize the Categorical Cross-Entropy Loss, using Adam as the optimizer with a fixed learning rate of  $10^{-3}$ . The metric selected to compare the inferred and the ground truth information is the *Unweighted Average Recall* (UAR). This metric allows us to account for the potential imbalance of the windowed sequences of data generated for the different activities. Hence, we define  $(1 - \text{UAR})$  as the validation error to monitor the training progress. Network parameters are updated in batches of 64 samples and trained during a maximum of 150 epochs. We implement an early stopping mechanism to stop training when the validation error does not improve for 20 consecutive epochs. To assess the models, we follow a nested *Leave-One-Subject-Out Cross-Validation* (LOSO-CV) approach, splitting the data in the inner loop into 5 participant-independent folds. Each fold in the inner loop is trained during a specific number of epochs. Therefore, when modeling all the training material in the outer loop and to prevent overfitting, the training epochs are determined by computing the median of the training epochs processed in each fold. The resulting model is tested on the initially excluded participant. In compliance with the LOSO-CV approach, we apply this routine recursively, so each participant in the dataset can be used to test the performance of the trained models.

## 5. EXPERIMENTAL RESULTS

This section summarizes the experiments performed in this work and analyzes the results obtained. The resting activity is

excluded from our experiments as, from a conceptual point of view, it can overlap with the lying, sitting, and standing activities. Nevertheless, the information collected during the resting activity is used in the context of our study to compute the personal, baseline heart rate (cf. Section 4.1.2). The pedometer information from 3 participants included in this first version of the harAGE dataset is corrupted. Consequently, we exclude all the data from these participants to train the models object of this study. We assess the performance of the models described in Section 4.2 from three different perspectives. Section 5.1 addresses the task as a binary classification problem. For this, we cluster the original activities into those that are static and those that are dynamic. We exclude the samples corresponding to the washing hands, strength workout, and flexibility workout activities. Section 5.2 tackles the recognition of the standard HAR dataset. In this case, we aim to model the lying, sitting, standing, walking, running, stairs climbing, and cycling activities and, therefore, we formulate the task as a 7-class classification problem. Finally, Section 5.3 addresses the task as a 10-class classification problem, targeting the automatic recognition of the whole set of activities considered in the harAGE dataset. Model performances are assessed by computing the UAR between the inferred and the ground truth annotations.

### 5.1. Binary Classification

The results obtained when tackling the task as a binary classification problem are summarized in **Table 2**. Analyzing the results, we observe that the multimodal models improve the performance of the unimodal models in most of the cases investigated. Comparing the performance of the multimodal models using the concatenation-based and the outer product-based approaches, the results indicate the suitability of the concatenation-based approach in this context, as it outperforms the outer product-based approach in 6 out of the 8 scenarios compared. When using the RNN-based architecture to extract deep learnt representations from the input modalities, the best UAR of 93.7% is obtained with the model exploiting the pedometer, the heart rate, and the accelerometer modalities fused using the concatenation-based approach. The highest UAR of 95.6% is achieved by the CNN-based architecture exploiting the three modalities together fused using the concatenation-based approach. The confusion matrix computed by comparing the activities inferred by this model and the ground truth annotations is depicted in **Figure 3**.

### 5.2. Standard HAR Classification

The results obtained when tackling the task as a 7-class classification problem are summarized in **Table 3**. The first observation of the results allows us to state that the multimodal models outperform the unimodal models in most of the cases investigated. The multimodal models using the RNN-based architecture to extract deep learnt representations from the input modalities and fusing the embedded information with the outer product-based approach surpass the models using the concatenation-based fusion in 3 out of the 4 scenarios compared. The multimodal models using the CNN-based architecture and

True label	Lying	151 (.680)	49 (.221)	12 (.054)	8 (.036)	—	—	1 (.005)	—	—	1 (.005)
	Sitting	38 (.186)	130 (.637)	30 (.147)	3 (.015)	1 (.005)	—	—	—	—	2 (.010)
	Standing	15 (.068)	29 (.131)	155 (.701)	13 (.059)	4 (.018)	—	1 (.005)	—	2 (.009)	2 (.009)
	Washing Hands	1 (.008)	5 (.040)	7 (.056)	101 (.815)	1 (.008)	2 (.016)	3 (.024)	1 (.008)	3 (.024)	—
	Walking	9 (.027)	18 (.055)	5 (.015)	2 (.006)	232 (.707)	16 (.049)	31 (.095)	—	5 (.015)	10 (.030)
	Running	1 (.004)	—	1 (.004)	3 (.011)	9 (.032)	240 (.860)	22 (.079)	1 (.004)	2 (.007)	—
	Stairs Climbing	—	—	—	3 (.010)	54 (.171)	51 (.162)	182 (.578)	12 (.038)	11 (.035)	2 (.006)
	Strength Workout	—	—	6 (.053)	6 (.053)	2 (.018)	1 (.009)	29 (.257)	39 (.345)	2 (.018)	28 (.248)
	Flexibility Workout	—	—	17 (.132)	15 (.116)	9 (.070)	3 (.023)	32 (.248)	7 (.054)	36 (.279)	10 (.078)
	Cycling	6 (.025)	20 (.085)	12 (.051)	—	4 (.017)	—	7 (.030)	12 (.051)	4 (.017)	171 (.725)
		Predicted label									
		Lying	Sitting	Standing	Washing Hands	Walking	Running	Stairs Climbing	Strength Workout	Flexibility Workout	Cycling

**FIGURE 5 |** Confusion matrix computed from the predictions inferred using the CNN-based approach exploiting  $f_{steps@hr@xyz}$  as input features. Each cell contains the absolute and the relative number of samples of the actual class (row) classified into each possible activity of the recognition set (column). The relative information is also depicted using a color-scale: dark colors illustrate high percentages, while light colors, low percentages. Empty cells indicate no samples from the actual activity are classified into the corresponding class.

fusing the embedded representations with the outer product-based approach improve the performance of the concatenation-based fusion in 4 out of the 4 cases compared. Although the  $f_{steps@hr@xyz}$  and the  $f_{steps@hr@xyz}$  models obtain the same mean from the individual UAR scores, the variance associated to the latter is lower. Hence, we consider the  $f_{steps@hr@xyz}$  model as the better of the two. The model using the RNN-based architecture with the highest UAR score of 62.5 % fuses the heart rate, and the accelerometer modalities with the outer product-based approach. The model with the highest UAR of 69.5 % implements the CNN-based architecture and fuses the pedometer, the heart rate, and the accelerometer modalities with the outer product-based approach. The confusion matrix computed by comparing the activities inferred by this model and the ground truth annotations is depicted in **Figure 4**.

### 5.3. Multi-Class harAGE Classification

The results obtained when tackling the task as a 10-class classification problem are summarized in **Table 4**. The results obtained indicate that the multimodal models surpass the unimodal models in most of the cases. From the results, we also observe that the multimodal models fusing the embedded

representations with the outer product-based approach outperform the concatenation-based fusion in all the cases investigated with one exception: the multimodal RNN-based network fusing the pedometer, heart rate, and accelerometer information using the concatenation-based approach surpasses the outer product-based fusion. The model with the best performance using the RNN-based architecture scores a UAR of 56.6 %, exploiting the pedometer, the heart rate, and the accelerometer modalities fused with the concatenation-based approach. The highest UAR of 60.8 % is obtained with the CNN-based model that fuses the pedometer, the heart rate, and the accelerometer modalities using the outer product-based approach. The confusion matrix computed by comparing the activities inferred by this model and the ground truth annotations is depicted in **Figure 5**. As it can be seen in the confusion matrix, the strength and the flexibility workout activities are the most difficult ones to be recognized and cause the highest confusion. While the samples corresponding to the strength workout activities tend to be misclassified into the stairs climbing and the cycling activities, the samples corresponding to the flexibility workout activities tend to be mainly misclassified into the stairs climbing activity.

## 6. CONCLUSIONS

This work focused on the use of an outer product-based approach to fuse the embedded representations learnt from the pedometer, the heart rate, and the accelerometer information collected using a smartwatch for the problem of HAR. The best results obtained when tackling the task as a 2-class, 7-class, and 10-class classification problem were achieved with the multimodal models using a CNN-based architecture to extract deep learnt representations from the pedometer, the heart rate, and the accelerometer modalities as input data. The outer product-based fusion obtained the highest UAR scores in the 7-class, and the 10-class problems, and ranked the second highest UAR score in the 2-class problem. These results supported the suitability of fusing the pedometer, the heart rate, and the accelerometer information with the proposed outer product-based approach for the task at hand.

The pre-processing applied to the accelerometer measurements is one of the limitations of this work. As described in Section 4.1.3, we computed the personal, debiasing parameters for the accelerometer measurements using all the data available from the current participant in the dataset. In a real-life scenario, these parameters should be computed and updated on the fly. In terms of model performance, we expect the trained models to underperform when a new user uses the system for the first time, and improve as the user keeps using the system, once the personal, debiasing accelerometer parameters stabilize.

Further research directions include the investigation of other techniques to fuse the information from the available modalities. Additionally, exploring whether the high performance of the binary classification problem can be used to improve the performance of the  $N$ -class classification problems—with  $N > 2$ —in a multi-task or a transfer learning set up might also be worth researching.

## REFERENCES

- Ahmed, N., Rafiq, J. I., and Islam, M. R. (2020). Enhanced human activity recognition based on smartphone sensor data using hybrid feature selection model. *Sensors* 20, 317. doi: 10.3390/s20010317
- Altun, K., and Barshan, B. (2010). "Human activity recognition using inertial/magnetic sensor units," in *Proceedings of the International Workshop on Human Behavior Understanding* (Istanbul: Springer), 38–51.
- Ashry, S., Ogawa, T., and Gomaa, W. (2020). CHARM-deep: continuous human activity recognition model based on deep neural network using IMU sensors of smartwatch. *Sensors* 20, 8757–8770. doi: 10.1109/JSEN.2020.2985374
- Bayat, A., Pomplun, M., and Tran, D. A. (2014). A study on human activity recognition using accelerometer data from smartphones. *Procedia Comput. Sci.* 34, 450–457. doi: 10.1016/j.procs.2014.07.009
- Chen, Y., and Shen, C. (2017). Performance analysis of smartphone-sensor behavior for human activity recognition. *IEEE Access* 5, 3095–3110. doi: 10.1109/ACCESS.2017.2676168
- Fox, K. R. (1999). The influence of physical activity on mental well-being. *Public Health Nutr.* 2, 411–418. doi: 10.1017/s1368980099000567
- Fox, S. M., and Naughton, J. P. (1972). Physical activity and the prevention of coronary heart disease. *Prevent. Med.* 1, 92–120.
- Hassan, M. M., Uddin, M. Z., Mohamed, A., and Almogren, A. (2018). A robust human activity recognition system using smartphone sensors and deep learning. *Future Gen. Comput. Syst.* 81, 307–313. doi: 10.1016/j.future.2017.11.029
- Holzinger, A., Malle, B., Saranti, A., and Pfeifer, B. (2021). Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Inf. Fusion* 71, 28–37. doi: 10.1016/j.inffus.2021.01.008
- Khaira, P., Kumar, P., and Imran, J. (2018). Combining CNN streams of RGB-D and skeletal data for human activity recognition. *Pattern Recognit. Lett.* 115, 107–116. doi: 10.1016/j.patrec.2018.04.035
- Khan, A. M., Lee, Y.-K., Lee, S. Y., and Kim, T.-S. (2010). "Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis," in *Proceedings of the 5th International Conference on Future Information Technology* (Busan: IEEE), 6.
- Kim, J., Koh, J., Kim, Y., Choi, J., Hwang, Y., and Choi, J. W. (2018). "Robust deep multi-modal learning based on gated information fusion network," in *Proceedings of the Asian Conference on Computer Vision* (Perth, WA: Springer), 90–106.
- Kwon, Y., Kang, K., and Bae, C. (2014). Unsupervised learning for human activity recognition using smartphone sensors. *Exp. Syst. Appl.* 41, 6067–6074. doi: 10.1016/j.eswa.2014.04.037
- Lara, O. D., Pérez, A. J., Labrador, M. A., and Posada, J. D. (2012). Centinela: a human activity recognition system based on acceleration and vital sign data. *Pervasive Mobile Comput.* 8, 717–729. doi: 10.1016/j.pmcj.2011.06.004

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institute of Computer Science, Foundation of Research and Technology—Hellas, Greece. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

AM-R, AS, MP, and BS conceptualized the study. AM-R ran the machine learning experiments. AS and MP engaged participants and collected data. AM-R and AS did literature analysis, manuscript preparation, and editing. All authors revised, read, and approved the final manuscript.

## FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

## ACKNOWLEDGMENTS

The authors would like to sincerely thank the participants who took part in the collection of the investigated dataset. We would also like to thank Georgios Athanassiou and Michalis Maniadakis for their contributions in the protocol design for collecting the data.

- Li, Q., Gravina, R., Li, Y., Alsamhi, S. H., Sun, F., and Fortino, G. (2020). Multi-user activity recognition: challenges and opportunities. *Inf. Fusion* 63, 121–135. doi: 10.1016/j.inffus.2020.06.004
- Lin, W.-Y., Verma, V. K., Lee, M.-Y., and Lai, C.-S. (2018). Activity monitoring with a wrist-worn, accelerometer-based device. *Micromachines* 9, 450. doi: 10.3390/mi9090450
- Mallol-Ragolta, A., Semertzidou, A., Pateraki, M., and Schuller, B. (2021). “harAGE: a novel multimodal smartwatch-based dataset for human activity recognition,” in *Proceedings of the 16th International Conference on Automatic Face and Gesture Recognition* (Jodhpur: IEEE), 7.
- Mekruksavanich, S., and Jitpattanakul, A. (2020). “Smartwatch-based human activity recognition using hybrid LSTM network,” in *Proceedings of Sensors* (Rotterdam: IEEE), 4.
- Penedo, F. J., and Dahn, J. R. (2005). Exercise and well-being: a review of mental and physical health benefits associated with physical activity. *Curr. Opin. Psychiatry* 18, 189–193. doi: 10.1097/00001504-200503000-00013
- Prakash, A., Chitta, K., and Geiger, A. (2021). “Multi-modal fusion transformer for end-to-end autonomous driving,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 7077–7087.
- Priyasad, D., Fernando, T., Denman, S., Sridharan, S., and Fookes, C. (2021). Memory based fusion for multi-modal deep learning. *Inf. Fusion* 67, 136–146. doi: 10.1016/j.inffus.2020.10.005
- Qi, J., Wang, Z., Lin, X., and Li, C. (2018). Learning complex spatio-temporal configurations of body joints for online activity recognition. *IEEE Trans. Hum. Mach. Syst.* 48, 637–647. doi: 10.1109/THMS.2018.2850301
- Ronao, C. A., and Cho, S.-B. (2014). “Human activity recognition using smartphone sensors with two-stage continuous hidden markov models,” in *Proceedings of the 10th International Conference on Natural Computation* (Xiamen: IEEE), 681–686.
- Ronao, C. A., and Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Exp. Syst. Appl.* 59, 235–244. doi: 10.1016/j.eswa.2016.04.032
- Shahmohammadi, F., Hosseini, A., King, C. E., and Sarrafzadeh, M. (2017). “Smartwatch based activity recognition using active learning,” in *Proceedings of the International Conference on Connected Health: Applications, Systems and Engineering Technologies* (Philadelphia, PA: IEEE), 321–329.
- Tapia, E. M., Intille, S. S., Haskell, W., Larson, K., Wright, J., King, A., et al. (2007). “Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor,” in *Proceedings of the 11th International Symposium on Wearable Computers* (Boston, MA: IEEE), 4.
- Wan, S., Qi, L., Xu, X., Tong, C., and Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. *Mobile Netw. Appl.* 25, 743–755. doi: 10.1007/s11036-019-01445-x
- Weiss, G. M., Timko, J. L., Gallagher, C. M., Yoneda, K., and Schreiber, A. J. (2016). “Smartwatch-based activity recognition: a machine learning approach,” in *Proceedings of the 3rd International Conference on Biomedical and Health Informatics* (Las Vegas, NV: IEEE), 426–429.
- Wu, A., and Han, Y. (2018). “Multi-modal circulant fusion for video-to-language and backward,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm: IJCAI), 1029–1035.
- Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Copenhagen: ACL), 1103–1114.
- Zhuang, Z., and Xue, Y. (2019). Sport-related human activity detection and recognition using a smartwatch. *Sensors* 19, 21. doi: 10.3390/s19225001

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Mallol-Ragolta, Semertzidou, Pateraki and Schuller. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Uniform Polylogarithmic Space Completeness

Flavio Ferrarotti<sup>1\*</sup>, Senén González<sup>1</sup>, Klaus-Dieter Schewe<sup>2</sup> and José María Turull-Torres<sup>3</sup>

<sup>1</sup> Software Competence Center Hagenberg, Hagenberg, Austria, <sup>2</sup> University of Illinois at Urbana-Champaign Institute (UIUC), Zhejiang University, Haining, China, <sup>3</sup> DIIT, Department of Engineering and Technological Research, Universidad Nacional de La Matanza, Buenos Aires, Argentina

It is well-known that polylogarithmic space (PolyL for short) does not have complete problems under logarithmic space many-one reductions. Thus, we propose an alternative notion of completeness inspired by the concept of uniformity studied in circuit complexity theory. We then prove the existence of a uniformly complete problem for PolyL under this new notion. Moreover, we provide evidence that uniformly complete problems can help us to understand the still unclear relationship between complexity classes such as PolyL and polynomial time.

**Keywords:** reductions, completeness, polylogarithmic space, PolyL, complexity theory

## OPEN ACCESS

### Edited by:

Daowen Qiu,  
Sun Yat-sen University, China

### Reviewed by:

Gokarna Sharma,  
Kent State University, United States  
Maria Alessandra Ragusa,  
University of Catania, Italy

### \*Correspondence:

Flavio Ferrarotti  
flavio.ferrarotti@scch.at

### Specialty section:

This article was submitted to  
Theoretical Computer Science,  
a section of the journal  
Frontiers in Computer Science

**Received:** 30 December 2021

**Accepted:** 16 March 2022

**Published:** 07 April 2022

### Citation:

Ferrarotti F, González S, Schewe K-D  
and Turull-Torres JM (2022) Uniform  
Polylogarithmic Space Completeness.  
Front. Comput. Sci. 4:845990.  
doi: 10.3389/fcomp.2022.845990

## 1. INTRODUCTION

The class of problems that can be decided by deterministic Turing machines using space bounded by a polynomial in the logarithm of the input size, is known in computational complexity as polylogarithmic space and usually denoted as PolyL (see e.g., Papadimitriou, 1994). Same as we know that every problem in L (logarithmic space) is in P (polynomial time), we have that every problem in PolyL is in QP (quasi-polynomial time). This latter class represents algorithms which run in sub-exponential time, more precisely in time bounded by  $2^{O(\log^c n)}$  for inputs of size  $n$  and some fixed constant  $c$ , and can thus be considered somehow more tractable than exponential time algorithms. Interestingly, the fastest known algorithm for checking graph isomorphism belongs to QP (see Babai, 2016).

It follows from the well-known space hierarchy theorem of Hartmanis et al. (1965) that the subset of problems in PolyL with space bounded above by  $\log^c n$  is strictly included in the subset bounded above by  $\log^{c+1} n$  for all integers  $c > 0$ . Therefore, PolyL cannot have a complete problem under logarithmic space many-one reductions. Note that if PolyL had such a complete problem  $A$ , then the space complexity of deciding this problem must be bounded above by  $\log^c n$  for some constant  $c$ , i.e., it must be in  $\text{DSPACE}(\log^c n)$  for some fixed  $c$ . If we now take a problem  $B$  in  $\text{DSPACE}(\log^{c+1} n) \setminus \text{DSPACE}(\log^c n)$ , we get (by our assumption on the completeness of  $A$ ) that there is a Turing machine that decides  $B$  in space  $O(\log^c n)$ , as  $B$  could be reduced to  $A$  using logarithmic space and then decided in  $O(\log^c n)$ . This means that  $B \in \text{DSPACE}(\log^c n)$ , contradicting that by the space hierarchy theorem  $\text{DSPACE}(\log^c n) \subset \text{DSPACE}(\log^{c+1} n)$ .

Even though PolyL does not have complete problems under logarithmic space many-one reductions, and thus can be considered less robust than L (its logarithmic counterpart), we show in this article that this perception can be challenged. Indeed, considering an alternative notion of completeness, we are able to show that it is still possible to isolate the most difficult problems in PolyL and to draw standard consequences of the kind entailed by the classical notion of completeness.

Our alternative notion of completeness (and hardness) is grounded in the concept of uniformity borrowed from circuit complexity theory (see Balcázar et al., 1990; Immerman, 1999; Murray and Williams, 2017 among others), hence we call it *uniform completeness*. A reviewer suggested that this is also related to Ragusa (2012). The intuitive idea is to consider a countably infinite family of problems instead of a single global problem. Each problem in the family corresponds to a fragment of a same global problem determined by a positive integer parameter. Such problem is uniformly complete for a given complexity class if there is a transducer Turing machine which given a positive integer as input builds a Turing machine in the required complexity class that decides the fragment of the problem corresponding to this parameter.

The article is organized as follows. In section 2, we introduce the necessary background from complexity theory and fix the notation used throughout the article. We dedicate section 3 to examine the quantified Boolean satisfiability problem and to define a restriction of this problem which, as we show in this article, captures the power and complexity of PolyL. We present the notion of uniform completeness (and hardness) in section 4. This constitutes the main novelty of this article. Using this new concept, we then prove in section 5 the existence of a uniform complete problem for PolyL. The problem is defined using the restriction of the quantified Boolean satisfiability problem introduced earlier in section 3. We present our conclusion in section 6.

## 2. PRELIMINARIES

We assume the reader is familiar with deterministic Turing machines and only include here the formal definitions that are required to fix the notation. We take these formal definitions from Balcázar et al. (1995).

**Definition 2.1.** A *deterministic Turing machine* with  $r$  tapes is a five-tuple  $M = \langle Q, \Sigma, \delta, q_0, F \rangle$ , where:

- $Q$  is the finite set of *states*.
- $\Sigma$  is the finite *tape alphabet*.
- $q_0 \in Q$  is the *initial state*.
- $F \subseteq Q$  is the *set of accepting final states*.
- $\delta : Q \times \Sigma^r \rightarrow Q \times \Sigma^{r-1} \times \{R, N, L\}^r$  is the (partial) *transition function* of  $M$ .

By the previous definition, there is one tape whose content cannot be changed by the transition function. This is the input tape, that we assume is read-only. Consequently, each configuration (a.k.a. instantaneous description or snapshot) does not need to include the contents of the input tape, as the position of the input head is enough to determine the current symbol read from the input.

**Definition 2.2.** A *configuration* of a Turing machine  $M$  with  $r$  tapes on a fixed input is a  $r+1$  tuple of the form:  $(q, i, w_2, \dots, w_r)$ , where  $q$  is the current state of  $M$ ,  $i$  is the position of the input tape head and each  $w_j \in \Sigma^* \# \Sigma^*$  represents the current content of the  $j$ -th work tape. The symbol  $\# \notin \Sigma$  marks that the tape head is reading the symbol immediately to its right. All symbols in the

infinite work tape  $j$  that do not appear in  $w_j$  are assumed to be the symbol blank “ $\sqcup$ ”. In the *initial configuration*  $(q_0, 0, \#, \dots, \#)$  of  $M$ , all work tapes are blank and the input tape head is scanning the leftmost cell. An *accepting configuration* is a configuration whose state is an accepting final state.

The concept of computation can now be defined as a sequence of configurations.

**Definition 2.3.** A *partial computation* of a Turing machine  $M$  with  $r$  tapes on an input string  $w = a_1, \dots, a_n$  is a (possibly infinite) sequence of configurations of  $M$ , in which each step from a configuration to the next is dictated by the transition function as follows: Assume a configuration *conf* with state  $q$ , input tape head in position  $i$  and each  $j$ -th work tape head scanning a corresponding symbol  $b_j$ . If  $\delta(q, a_i, b_2, \dots, b_r) = (q', b'_2, \dots, b'_r, m_1, \dots, m_r)$ , then the state in the next configuration is  $q'$ , the position of the input tape head  $i'$  equals  $i+1$ ,  $i-1$  or  $i$  depending on whether  $m_1$  is  $R$ ,  $L$ , or  $N$ , respectively, and each string  $w'_j$  ( $2 \leq j \leq r$ ) representing the contents of the  $j$ -th work tape equals the corresponding string  $w_j$  in *conf* with the possible exceptions of the symbol  $b_j$  in the position immediately to the right of  $\#$  in  $w_j$  and the position of  $\#$  itself. The former is replaced in  $w'_j$  by  $b'_j$ . The latter is moved one position to the right if  $m_j = R$ , to the left if  $m_j = L$ , or not moved if  $m_j = N$ . If the  $\#$  in  $w_j$  is in the leftmost position and  $m_j = L$ , then  $\#$  remains in the same place in  $w'_j$ . Likewise, if  $i = 0$  and  $m_1 = L$ , then the head of the input tape remains in place, i.e.,  $i' = i$ . A *computation* is a partial computation that starts with the initial configuration of  $M$ , and ends in a configuration where no more steps can be performed.

If a computation of a Turing machine  $M$  ends in an accepting configuration, then we call it an *accepting computation* of  $M$ . In this case, the word in the input tape is *accepted* by  $M$ . The *language*  $L(M)$  accepted by  $M$  is the set of all words accepted by  $M$ .

Non-deterministic Turing machines simply relax the definition of transition function  $\delta$  with respect to their deterministic counterparts, so that every move is *not* necessarily determined uniquely by the current configuration.

**Definition 2.4.** A *non-deterministic Turing machine* with  $r$  tapes is a five-tuple  $M = \langle Q, \Sigma, \delta, q_0, F \rangle$ , where  $Q$ ,  $\Sigma$ ,  $q_0$  and  $F$  are exactly as in Definition 2.1 and the transition function is defined by

$$\delta : Q \times \Sigma^r \rightarrow \mathcal{P}(Q \times \Sigma^{r-1} \times \{R, N, L\}^r)$$

where  $\mathcal{P}(A)$  denotes the powerset of  $A$ .

The previous definitions of configurations and computations apply in exactly the same way to non-deterministic Turing machines. However, they can have more than one computation for a given input. Thus, a non-deterministic Turing machine  $M$  *accepts* an input string  $w$  iff there exists a computation of  $M$  on  $w$  ending in an accepting configuration.

Complexity theory mainly concerns the classification of computational problems in terms of required Turing machine

resources, and the study of the relationship between the resulting classes.

**Definition 2.5.** Let  $n$  denote the size of a Turing machine input, i.e., the number of non-blank cells in the input tape, and let  $t$  and  $s$  be functions such that  $t(n) > n + 1$  and  $s(n) \geq 1$ . We define  $DTIME(t)$ ,  $NTIME(t)$  as the classes of all languages accepted by deterministic and non-deterministic Turing machines, respectively, whose running time is bounded above by  $t(n)$ . Similarly, We define  $DSPACE(s)$ ,  $NSPACE(s)$  as the classes of all languages accepted by deterministic and non-deterministic Turing machines, respectively, whose work space is bounded above by  $s(n)$ . Running time means the number of transitions in a computation. Work space is the number of different work tape cells (counting all work tapes) used during a computation.

In this work we concentrate in the class of languages accepted by deterministic Turing machines with polylogarithmic bounded space, i.e., in the class known as PolyL. Of course, we also need to reference some related complexity classes. They are formally defined as follows.

**Definition 2.6.** Let  $\log n$  denote the logarithm base 2 of  $n$ , i.e.,  $\log_2 n$ , and  $\log^k n$  denote  $(\lceil \log n \rceil)^k$ , where  $\lceil \log n \rceil$  is the least integer greater than or equal to  $\log n$ .

$$L = \bigcup_{c \geq 1} DSPACE(c \cdot \log n); \quad NL = \bigcup_{c \geq 1} NSPACE(c \cdot \log n); \quad \text{PolyL} = \bigcup_{k \geq 0} DSPACE(\log^k n)$$

$$P = \bigcup_{k \geq 0} DTIME(n^k); \quad NP = \bigcup_{k \geq 0} NTIME(n^k); \quad PSPACE = \bigcup_{k \geq 0} DSPACE(n^k)$$

It is well-known that each deterministic class is closed under complementation and that each deterministic class is included in its non-deterministic counterpart, but it is not known whether this inclusion is strict. It is also well-known that NL is closed under complementation, included in P and strictly included in PSPACE. Also, NP is included in PSPACE. All these results can be found in classical complexity theory books such as in Papadimitriou (1994) and in Balcázar et al. (1995). Obviously, NL is included in PolyL, but it is unclear how PolyL compares with P. We only know that  $\text{PolyL} \neq P$ , since P has a complete problem under logarithmic space many-one reductions but polyL does not due to the space hierarchy theorem. It is not expected for polyL to be strictly contained in P. Whether the converse is true, it is also unclear.

One of the most important sort of intuitions about complexity classes and their interrelationships is provided by the concepts of reducibility and complete problem. We again borrow the definitions of these well-known concepts from Balcázar et al. (1995).

**Definition 2.7.** A language  $A$  is *polynomial time many-one reducible* to a language  $B$  iff there is a function  $f: \Sigma^* \rightarrow \Sigma^*$ , computable in polynomial time by a transducer Turing machine, such that  $w \in A$  iff  $f(w) \in B$  for all  $w \in \Sigma^*$ .

As usual, we denote that  $A$  is reducible to  $B$  by  $A \leq_m B$ , and if  $f$  is the function that defined this reduction, then we say that  $A \leq_m B$  via  $f$ . Polynomial time many-one reductions are sometimes called Karp reductions. For the classes P, L and NL, Karp reductions are considered too strong, since they have complete problems via logarithmic space reductions. The concept of completeness provides important insight about the most difficult problems inside a complexity class.

**Definition 2.8.** Given a complexity class  $C$ ,

- A language  $A$  is  $C$ -hard iff for any language  $B$  in  $C$ , we have that  $B \leq_m A$ .
- A language  $A$  is  $C$ -complete iff it is  $C$ -hard and  $A \in C$ .

### 3. A RESTRICTED QUANTIFIED BOOLEAN SATISFIABILITY PROBLEM

Our aim is to define a problem that captures the power and complexity of PolyL. We start from a well-known problem that captures these features in another deterministic space complexity class, namely the PSPACE-complete problem of determining the satisfiability of quantified Boolean sentences (QSAT for short), and explore a restriction of this problem so that it can be uniformly solved in PolyL.

**Definition 3.1.** Let  $V$  be a set of Boolean variables, i.e., a set of symbols that can take the value 0 (false) or 1 (true). The class of *Boolean formulae* over  $V$  is defined by the following rules:

- Constants 0 and 1 are Boolean formulae.
- If  $x \in V$ , then  $x$  is a Boolean formulae.
- If  $\varphi$  and  $\psi$  are Boolean formulae, then  $(\varphi \vee \psi)$ ,  $(\varphi \wedge \psi)$ , and  $\neg(\varphi)$  are Boolean formulae.
- Nothing else is a Boolean formulae.

The class of *quantified Boolean formulae* is the smallest class defined by the rules:

- Every Boolean formula is a quantified Boolean formula.
- If  $x \in V$  and  $\varphi$  is a quantified Boolean formula, then  $\exists x \varphi$  and  $\forall x \varphi$  are quantified Boolean formulae.

Notice that the previous definition implies that a quantified Boolean formula is always in prenex normal form, i.e., every quantified Boolean formula consists of a (possible empty) prefix of quantifiers followed by a quantifier-free Boolean formula. This is of course not necessary, but it is convenient. We also assume w.l.o.g. that every variable that appears in a formula is quantified in the prefix at most once.

Let  $\varphi[x/a]$ , where  $a \in \{0, 1\}$  and  $\varphi$  is a quantified Boolean formula, denote the formula obtained by substituting  $a$  for every occurrence of  $x$  in  $\varphi$ . The semantics of quantified Boolean formulae can be formally defined as follows.

**Definition 3.2.** Let  $v: X \rightarrow \{0, 1\}$  be a Boolean assignment and  $\varphi$  a quantified Boolean formula, the *truth value of  $\varphi$  under  $v$*  is defined recursively by the rules:

- $v(\varphi) = 0$  if  $\varphi = 0$ .
- $v(\varphi) = 1$  if  $\varphi = 1$ .
- $v(\varphi) = v(x)$  if  $\varphi = x$  for some  $x \in V$ .
- $v(\varphi) = v(\psi) + v(\alpha)$  (Boolean addition) if  $\varphi = (\psi \vee \alpha)$ .
- $v(\varphi) = v(\psi) \cdot v(\alpha)$  (Boolean multiplication) if  $\varphi = (\psi \wedge \alpha)$ .
- $v(\varphi) = \overline{v(\psi)}$  (Boolean complement) if  $\varphi = \neg(\psi)$ .
- $v(\varphi) = v(\psi[x/0]) + v(\psi[x/1])$  (Boolean addition) if  $\varphi = \exists x\varphi$ .
- $v(\varphi) = v(\psi[x/0]) \cdot v(\psi[x/1])$  (Boolean multiplication) if  $\varphi = \forall x\varphi$ .

Boolean formulae can be encoded as words over a finite alphabet and thus written down in Turing machine tapes. Here, we encode quantified Boolean formulae as words over the following alphabet:

$$\Sigma_{\text{QBF}} = \{\wedge, \vee, \neg, \forall, \exists, (, ), 0, 1, \text{true}, \text{false}\}$$

where each variable is represented by the binary expression of its subindex, and *true* and *false* denote the Boolean constants.

We can now formally define the QSAT problem as well as its restriction for PolyL.

**Definition 3.3.** Let QBF denote the set of quantified Boolean formulae encoded as words of a fixed alphabet  $\Sigma$  and let  $\text{var}(\varphi)$  for  $\varphi \in \text{QBF}$  denote the set of variables encoded in  $\varphi$ .

- QSAT is the subset of QBF formed by all encodings of quantified Boolean sentences (i.e., formulae without free variables) that evaluate to “true”.
- $\text{QSAT}_k^{\text{PL}} = \{\varphi \in \text{QSAT} \mid |\text{var}(\varphi)|^3 \leq \log^k |\varphi|\}$ .

It is well-known that QSAT is complete for PSPACE under Karp reductions. See for instance Theorem 3.29 and its associated lemmata in Balcázar et al. (1995). Using a similar strategy plus a new concept of uniform completeness defined in the next section, we show in this article that the family of problems  $\{\text{QSAT}_k^{\text{PL}}\}_{k \geq 0}$  captures the essence of the most difficult problems in PolyL.

## 4. UNIFORM COMPLETENESS

The new notion of completeness (and hardness) that we define in this section is grounded in the concept of uniformity borrowed from circuit complexity theory (see e.g., Balcázar et al., 1990; Immerman, 1999), hence we call it *uniform completeness*.

As a first step we need to define a notion of uniformity for Turing machines.

**Definition 4.1.** Let  $\mathcal{M}$  be a countably infinite class of Turing machines such that for every integer  $k > 0$  there is exactly one machine  $M_k \in \mathcal{M}$ . We say that  $\mathcal{M}$  is *uniform* if there is a Turing machine  $M_{\mathcal{M}}$  which for every input  $k \geq 0$  builds an encoding of the corresponding  $M_k \in \mathcal{M}$ .

Instead of looking at isolated languages (or problems), we are concerned here with families of languages (or problems).

**Definition 4.2.** A *language family*  $\mathcal{L}$  is a countably infinite class of languages of a same finite vocabulary. We say that  $\mathcal{L}$  is *compatible* with a language  $A$  if  $\bigcup_{L_i \in \mathcal{L}} L_i = A$ .

Consequent with the previous definitions, we now consider decidability in the context of families of languages and uniform classes of Turing machines.

**Definition 4.3.** Let  $\mathcal{L}$  be a language family and  $\mathcal{M}$  be an uniform class of Turing machines. We say that  $\mathcal{M}$  *uniformly decides*  $\mathcal{L}$  if for every  $L_i \in \mathcal{L}$  there is an  $M_j \in \mathcal{M}$  such that  $M_j$  decides  $L_i$ .

The following definition clarifies when we can say in this context that a language (uniformly) belongs to a complexity class.

**Definition 4.4.** Let  $\mathcal{C}$  be a complexity class and  $\mathcal{L}$  be a language family. A language  $A$  is *uniformly in  $\mathcal{C}$  via  $\mathcal{L}$*  if the following holds:

- $\mathcal{L}$  is compatible with  $A$ .
- There is a uniform class of Turing machines  $\mathcal{M}$  which uniformly decides  $\mathcal{L}$ .
- Each Turing machine in  $\mathcal{M}$  satisfies the same resource restrictions that define  $\mathcal{C}$ .

For a uniform reduction of a language to a language family, we simply require the existence of a standard polynomial time many-one reduction to a single member of that family.

**Definition 4.5.** A language  $A$  is *uniformly many-one reducible* to a language family  $\mathcal{L}$  (denoted  $A \leq_m^u \mathcal{L}$ ) iff there is a language  $L \in \mathcal{L}$  such that  $A$  is polynomial time many-one reducible to  $L$ , i.e., iff  $A \leq_m L$ .

We have now the necessary tools to define our uniform notion of completeness (and hardness).

**Definition 4.6.** Given a complexity class  $\mathcal{C}$  and a language family  $\mathcal{L}$ ,

- A language  $A$  is *uniformly  $\mathcal{C}$ -hard via  $\mathcal{L}$*  iff  $\mathcal{L}$  is compatible with  $A$  and for any language  $B$  in  $\mathcal{C}$ , we have that  $B \leq_m^u \mathcal{L}$ .
- A language  $A$  is *uniformly  $\mathcal{C}$ -complete via  $\mathcal{L}$*  iff it is uniformly  $\mathcal{C}$ -hard via  $\mathcal{L}$  and uniformly in  $\mathcal{C}$  via  $\mathcal{L}$ .

Classical complete problems in complexity theory lead to some interesting consequences such as Corollary 3.19c in Balcázar et al. (1995) which states that if a PSPACE-complete problem under Karp reductions is in P, then  $\text{PSPACE} = \text{P}$ . The following lemma shows that our somehow “weaker” notion of uniform completeness still allows us to derive similar kinds of results.

**Lemma 4.1.** Let  $A$  be uniformly PolyL-complete via a problem family  $\mathcal{L}$ . If  $A$  is uniformly in P via  $\mathcal{L}$ , then  $\text{PolyL} \subseteq \text{P}$ .

*Proof:* Let  $\mathcal{M}$  and  $\mathcal{M}'$  be classes of deterministic Turing machines that uniformly decide  $\mathcal{L}$  and, respectively, witness that  $A$  is uniformly in PolyL and P. Since we assume that  $A$  is uniformly complete for PolyL via  $\mathcal{L}$ , it follows by definition that for each language  $B$  in PolyL there is an  $L$  in  $\mathcal{L}$  such that  $B \leq_m L$ , and thus also a corresponding Turing machine  $M_{B,L}$ .



that computes this reduction in polynomial time. Furthermore, the fact that  $A$  is uniformly in  $P$  via  $\mathcal{L}$  implies (again by definition) that there is a deterministic Turing machine  $M_L \in \mathcal{M}'$  that decides  $L$  in polynomial time. Therefore, we can define a deterministic Turing machine  $M_B$  that decides  $B$  in polynomial time. This shows that, under the assumptions in this lemma,  $\text{PolyL} \subseteq P$ . The machine  $M_B$  works by simply assembling together  $M_{B,L}$  and  $M_L$ , redirecting the output of  $M_{B,L}$  to a work tape and making  $M_L$  read its input from that work tape.  $\square$

It is interesting to note that  $\text{PolyL}$  is included in the class of problems that have quasi-polynomial time algorithms. This class is defined as  $\text{QP} = \bigcup_{k \geq 0} \text{DTIME}(2^{\log^k n})$  (see Babai, 2016 among others).

## 5. A UNIFORMLY COMPLETE LANGUAGE

In this section we show that the language  $\text{QSAT}^{pl} = \bigcup_{k \geq 0} \text{QSAT}_k^{pl}$  captures the essence of the most difficult problems in  $\text{PolyL}$ . That is, we prove the following result.

**Theorem 5.1.**  *$\text{QSAT}^{pl}$  is uniformly  $\text{PolyL}$ -complete via the language family  $\mathcal{L} = \{\text{QSAT}_k^{pl}\}_{k \geq 0}$ .*

As in classical complexity theory, we essentially need to prove that  $\text{QSAT}^{pl}$  is  $\text{PolyL}$ -hard and that it is indeed in this class. The subtle but important difference lies in the fact this is not possible in the traditional sense. We use instead the concept of uniform completeness (and hardness) introduced in Definition 4.6. Theorem 5.1 is thus a direct consequence of the fact that the following two conditions hold *via* the language family  $\mathcal{L} = \{\text{QSAT}_k^{pl}\}_{k \geq 0}$ :

- The language  $\text{QSAT}^{pl}$  is uniformly  $\text{PolyL}$ -hard.
- The language  $\text{QSAT}^{pl}$  is uniformly in  $\text{PolyL}$ .

Lemma 5.2 and 5.3 below prove, respectively, that both conditions are met.

**Lemma 5.2.**  *$\text{QSAT}^{pl}$  is uniformly  $\text{PolyL}$ -hard via  $\mathcal{L}$ .*

*Proof:* Since by definition  $\mathcal{L} = \{\text{QSAT}_k^{pl}\}_{k \geq 0}$  and  $\text{QSAT}^{pl} = \bigcup_{k \geq 0} \text{QSAT}_k^{pl}$ , it is trivial to see that  $\mathcal{L}$  is compatible with  $\text{QSAT}^{pl}$ .

Now we need to show that for any language  $B$  in  $\text{PolyL}$ ,  $B \leq_m^u \mathcal{L}$ . Since this is the case if there is a language  $\text{QBF}_k^{pl}$  for some  $k$  such that  $B \leq_m \text{QBF}_k^{pl}$ , we only need to show that we can build a quantified Boolean formula  $\varphi$  from the specification of a Turing machine  $M$  with polylogarithmic space bound and input  $w$ , such that  $\varphi$  evaluates to 1 iff  $M$  accepts  $w$ . Notice that the  $k$  can be as big as necessary and that there will always be a  $k$  big enough so that the encoding of  $\varphi$  belongs to  $\text{QBF}_k^{pl}$ . Thus the formula  $\varphi$  can be built exactly as in the proof that  $\text{QSAT}$  is  $\text{PSPACE}$ -hard (see for instance Theorem 3.29 in Balcázar et al., 1995), since  $\text{PolyL}$  is included in  $\text{PSPACE}$ .  $\square$

**Lemma 5.3.**  *$\text{QSAT}^{pl}$  is uniformly in  $\text{PolyL}$  via  $\mathcal{L}$ .*

*Proof:* We have already seen in the proof of the previous lemma that  $\mathcal{L}$  is compatible with  $\text{QSAT}^{pl}$ . Thus, we need to show that there is a uniform class of Turing machines  $\mathcal{M}$  which uniformly decides  $\mathcal{L}$  and that each Turing machine in  $\mathcal{M}$  works in polylogarithmic space.

We start by showing that, for every  $k \geq 0$ , the language  $\text{QSAT}_k^{pl}$  is in  $\text{DSPACE}(\log^k n)$ . Let  $\Sigma_{\text{QBF}} = \{\wedge, \vee, \neg, \forall, \exists, (, ), 0, 1, \text{true}, \text{false}\}$ . As discussed in section 3, we can encode arbitrary quantified Boolean formulae as words over this finite alphabet. We build a deterministic Turing machine  $M_k$  that takes as input a word  $w \in \Sigma_{\text{QBF}}^*$  which encodes a (not necessarily well-formed) quantified Boolean formula  $\varphi$  and decides whether  $w \in \text{QSAT}_k^{pl}$  working in space bounded above by  $\log^k |w|$ .

Let *eval* be the recursive procedure described in **Algorithm 1** which computes the value of a quantified Boolean formula in prenex normal form. If the length of  $w$  is  $n$ , then it is clear that the depth of the recursion defining *eval* cannot exceed this number, since the number of variables must be less than  $n$ . Furthermore, since we actually need to decide whether  $w \in \text{QSAT}_k^{pl}$ , we can stop the recursion and return false if the quantifier free part of the formula has not been reached at a recursion depth of  $|\text{var}(\varphi)|$  which by definition of  $\text{QSAT}_k^{pl}$  is less than  $\log^k n$ .

To implement *eval* we can use a stack, where in each entry we record the quantifier prefix up to that point, using a four-tuple of the form  $(Q_i, \bar{b}, v_1, v_2)$  for each quantifier  $Q_i$  in the prefix of the formula. The components of this tuple are as follows:  $Q_i$  is either  $\forall$  or  $\exists$ ,  $\bar{b}$  is the index in binary of the quantified variable  $x_i$ ,  $v_1$  is the truth value 0 or 1 assigned to this variable (initially 0) and  $v_2$  records the truth value of the sub-formulae  $\psi_i$  in  $Q_i x_i \psi_i$ . The value of  $v_2$  is blank if the sub-formula  $\psi_i$  has not been evaluated yet. Once  $\psi_i$  has been evaluated for first time with  $x_i = 0$ , the returned truth value 0 or 1 is stored in  $v_2$ ,  $v_1$  is updated to the value 1 and the subformula  $\psi_i$  is evaluated again with  $x_i = 1$ . At this point, we update  $v_2$  to the truth value obtained by taking the disjunction or conjunction of its current value with the one returned by  $\psi_i$ , depending on whether  $Q_i$  is  $\exists$  or  $\forall$ , respectively. This value  $v_2$  is then returned as value for the corresponding call *eval*( $Q_i x_i \psi_i$ ).

Given that the described approach needs space  $|\text{var}(\varphi)| \cdot (8 + \log |\text{var}(\varphi)|)$  for each stack entry, and that we have seen that the maximum recursion depth needed in our case is  $|\text{var}(\varphi)|$ , we get

---

**Algorithm 1** Evaluation of a quantified Boolean formula in prenex normal form.

---

- 1: **procedure** *eval*( $\varphi$ )
- 2:   **if**  $\varphi$  is quantifier-free **then**
- 3:     **return** *eval\_quantifier\_free*( $\varphi$ )
- 4:   **if**  $\varphi$  has the form  $\forall x \psi$  **then**
- 5:     **return** *eval*( $\psi[x/0]$ )  $\wedge$  *eval*( $\psi[x/1]$ )
- 6:   **if**  $\varphi$  has the form  $\exists x \psi$  **then**
- 7:     **return** *eval*( $\psi[x/0]$ )  $\vee$  *eval*( $\psi[x/1]$ )

that working space bounded by  $|var(\varphi)|^3$  is enough to implement this evaluation strategy for the quantifier prefix of  $\varphi$ .

Regarding the evaluation of the quantifier free part of  $\varphi$ , note that every time that we reach the last quantifier in the prefix, we have a full valuation for the variables in the quantifier free subformula. Thus we can evaluate this quantifier free subformula in space bounded by  $\log n$ . Note that the algorithm in Buss (1987) for the evaluation of Boolean formulas with variables and a value assignment works in alternating logarithmic time, which is known to be in L (i.e., in deterministic logarithmic space). See Theorem 2.32 in Immerman (1999) among other sources.

Thus, the size of the stack is what determines the upper bound in the space needed by  $M_k$  to decide whether  $w$ , i.e., the encoding of  $\varphi$ , is in  $QSAT_k^{pl}$ . Since this size is  $|var(\varphi)|^3$  and by definition of  $QSAT_k^{pl}$  we know that  $|var(\varphi)|^3 \leq \log^k n$ , we get that  $M_k$  can decide whether  $w \in QSAT_k^{pl}$  using space bounded above by  $\log^k |w|$ .

Clearly, the class  $\mathcal{M} = \bigcup_{k \geq 0} M_k$ , where each  $M_k$  is as described above, uniformly decides the language  $QSAT^{pl}$ . Furthermore, since we define  $M_k$  constructively, there is a Turing machine  $M_{\mathcal{M}}$  which for every input  $k$  builds an encoding of the corresponding  $M_k \in \mathcal{M}$ . This concludes our proof.  $\square$

## 6. CONCLUSION

In this article, we explore an alternative notion of completeness for PolyL. This notion is inspired by the concept of uniformity from circuit complexity theory. This results in a new concept of uniform completeness, which shows that we can still isolate the most difficult problems inside PolyL and draw some of the usual interesting conclusions entailed by the classical notion of complete problem (see in particular Lemma 4.1). The result is

relevant for it has been well-known since a long time that PolyL has no complete problems in the usual sense. It is plausible that this new concept of uniform completeness can be applied to other interesting complexity classes for which there are no (known) complete problems. The hope is to further our understanding of practically relevant complexity classes. Examples of such classes are deterministic and non-deterministic polylogarithmic time (see Ferrarotti et al., 2020 and Ferrarotti et al., 2021 among others) as well as the well-known quasi-polynomial time.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## FUNDING

The research reported in this article has been partially funded by the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Federal Ministry for Digital and Economic Affairs (BMDW), and the Province of Upper Austria in the frame of the COMET - Competence Centers for Excellent Technologies Programme managed by the Austrian Research Promotion Agency FFG (Österreichische Forschungsförderungsgesellschaft FFGNr. 865891).

## REFERENCES

- Babai, L. (2016). "Graph isomorphism in quasipolynomial time," in *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing* (Cambridge, MA), 684–697.
- Balcázar, J. L., Díaz, J., and Gabarró, J. (1990). *Structural Complexity II*, Vol. 22 of *EATCS Monographs on Theoretical Computer Science*. Berlin; Heidelberg; New York, NY; London; Paris; Tokyo; Hong Kong: Springer.
- Balcázar, J. L., Díaz, J., and Gabarró, J. (1995). *Structural Complexity I*, 2nd Edn. Berlin; Heidelberg; New York, NY; London; Paris; Tokyo; Hong Kong: Springer.
- Buss, S. R. (1987). "The boolean formula value problem is in ALOGTIME," in *Proceedings of the 19th Annual ACM Symposium on Theory of Computing*, ed A. V. Aho (New York, NY: ACM), 123–131.
- Ferrarotti, F., Gonzales, S., Schewe, K., and Turull Torres, J. M. (2020). A restricted second-order logic for non-deterministic poly-logarithmic time. *Log. J. IGPL* 28, 389–412. doi: 10.1093/jigpal/jzz078
- Ferrarotti, F., González, S., Turull Torres, J. M., Van den Bussche, J., and Virtema, J. (2021). Descriptive complexity of deterministic polylogarithmic time and space. *J. Comput. Syst. Sci.* 119, 145–163. doi: 10.1016/j.jcss.2021.02.003
- Hartmanis, J., Lewis, P. M., and Stearns, R. E. (1965). "Hierarchies of memory limited computations," in *6th Annual Symposium on Switching Circuit Theory and Logical Design* (Ann Arbor, MI: IEEE), 179–190.
- Immerman, N. (1999). *Descriptive Complexity*. New York, NY: Springer.
- Murray, C. D. and Williams, R. R. (2017). On the (non) np-hardness of computing circuit complexity. *Theory Comput.* 13, 1–22. doi: 10.4086/toc.2017.v013a004
- Papadimitriou, C. H. (1994). *Computational Complexity*. Reading, MA; Menlo Park, CA; New York, NY; Don Mills, ON; Wokingham; Amsterdam; Bonn; Sydney, NSW; Singapore; Tokyo; Madrid; Milan; Paris: Addison-Wesley.
- Ragusa, M. A. (2012). Parabolic herz spaces and their applications. *Appl. Math. Lett.* 25, 1270–1273. doi: 10.1016/j.aml.2011.11.022

**Conflict of Interest:** FF and SG were employed by Software Competence Center Hagenberg.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ferrarotti, González, Schewe and Turull-Torres. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Improving Mobile Device Security by Embodying and Co-adapting a Behavioral Biometric Interface

Avinash Jairam<sup>1</sup>, Tzipora Halevi<sup>1,2</sup> and Theodore Raphan<sup>1,2,3\*</sup>

<sup>1</sup> Ph.D. Program in Computer Science, Graduate Center of CUNY, New York, NY, United States, <sup>2</sup> Department of Computer and Information Science, Brooklyn College of CUNY, Brooklyn, NY, United States, <sup>3</sup> Ph.D. Program in Psychology and Neuroscience, Graduate Center of CUNY, New York, NY, United States

## OPEN ACCESS

### Edited by:

Andrej Košir,  
University of Ljubljana, Slovenia

### Reviewed by:

Robertas Damasevicius,  
Silesian University of  
Technology, Poland  
Mohammad Javed Morshed  
Chowdhury,  
La Trobe University, Australia

### \*Correspondence:

Theodore Raphan  
raphan@nsi.brooklyn.cuny.edu

### Specialty section:

This article was submitted to  
Computer Security,  
a section of the journal  
Frontiers in Computer Science

**Received:** 15 August 2021

**Accepted:** 07 March 2022

**Published:** 11 April 2022

### Citation:

Jairam A, Halevi T and Raphan T  
(2022) Improving Mobile Device  
Security by Embodying and  
Co-adapting a Behavioral Biometric  
Interface.  
Front. Comput. Sci. 4:754716.  
doi: 10.3389/fcomp.2022.754716

At present, interfaces between users and smart devices such as smart phones rely primarily on passwords. This has allowed for the intrusion and perturbation of the interface between the user and the device and has compromised security. Recently, Frank et al. have suggested that security could be improved by having an interface with biometric features of finger swiping. This approach has been termed touchalytics, in maintaining cybersecurity. The number of features of finger swiping have been large (32) and have been made available as a public database, which we utilize in our study. However, it has not been shown which of these features uniquely identify a particular user. In this paper, we study whether a subset of features that embody human cognitive motor features can be used to identify a particular user. We consider how the security might be made more efficient embodying Principal Component Analysis (PCA) into the interface, which has the potential of reducing the features utilized in the identification of intruders. We compare the accuracy and performance of the reduced feature space to that of having all the features. Embodying a robust continuous authentication system will give users an extra layer of security and an increased sense of peace of mind if their devices are lost or stolen. Consequently, such improvements may prevent access to sensitive information and thus will save businesses money. Consequently, such improvements may prevent access to sensitive information and thus will save businesses money. If continuous authentication models become successful and easily implementable, embodiment and co-adaptation of user authentication would inhibit the growing problem of mobile device theft.

**Keywords:** Behavioral Biometrics, computer security, keystroke dynamics, machine learning, touchalytics, embodiment, co-adaptation

## INTRODUCTION

Communication between individuals has an inherent authentication problem to determine if the sender is who he/she claims to be. This authentication problem has intrigued mankind for millennia. One of the first authentication systems dealt with the problem of point-to-point communication and addressed whether we could trust the person on the other end of a conversation (Dooley, 2013). As society expanded, humans have developed increasingly better systems for communicating over long distances. With each advancement, there exist unscrupulous individuals who will exploit weaknesses in the communication link to prey upon and exploit

unsuspecting users (Dooley, 2013). In fact, the wider the communication links and the greater the distance they span, the more likely that the communications medium will become a target for impersonators. Thus, at each stage of advancement, the need for verifying the identity of individuals far away and the need for better authentication systems have become increasingly important. One useful approach was the development of what has been referred to as biometric authentication (Bhattacharyya et al., 2009).

The idea of using biometric features, such as retinal scanning, gait characteristics, EEG biometrics, ear biometrics or combinations thereof have been a subject of investigation for quite some time in order to identify an individual's personal identity (Horst et al., 2019; Ma et al., 2020; Olanrewaju et al., 2020). These studies are based on the idea that every human has a unique anatomy, resulting in unique behavioral characteristics. For example, each human has a unique vocal tone because of the size and shape of the mouth and throat (Bhattacharyya et al., 2009). However, these techniques have described general characteristics of human features. For example, gait characteristic such as stride length, step frequency, and gait velocity, which are features of human gait (Cho et al., 2006, 2010; Osaki et al., 2007, 2008; Raphan, 2020) are difficult to implement for continuous monitoring by mobile handheld devices, which are generally used to examine documents and may be used while sitting or using transportation without walking. Neither is EEG an adequate medium for implementing continuous authentication of a user when these mobile devices are used in a wide range of environments. Finger swiping on the other hand has become the natural interface between humans and mobile devices and can be used in a wide range of environments. Therefore a concerted effort has been made to understand how the biometrics of finger swiping can be incorporated into the user interface of mobile devices (Frank et al., 2013).

The evolution of touch dynamics has progressed from the time of the introduction of the telegraph system to the present (Figure 1). The advent of the telegraph allowed messages to travel across the Atlantic within hours. Accordingly, the telegraph quickly became the information superhighway of its era and governments around the world soon adapted this technology (Telegraph, 2016). To ensure secure transmission of data, telegraph operators realized that they could identify each other by the timing patterns of their fellow operators over the medium. That is, the character set of Morse Code is the language of the telegraph and it consists of a series of dots and dashes. Telegraph operators noticed that the timing interval between when a dot and a dash is transmitted appeared to be unique to the user on the other end. Therefore, a pair of telegraph operators can become familiar with their counterpart's timing intervals and thus could identify each other and authenticate the transmission (Jenkins et al., 2011).

During World War II, the American military intelligence community developed the "FIST," which attempted to improve the security of Morse Code over wireless networks (Jenkins et al., 2011). Rapidly moving army units needed to communicate with their rear echelon commanders in real time. However, the dynamic nature of a battlefield made it impossible to

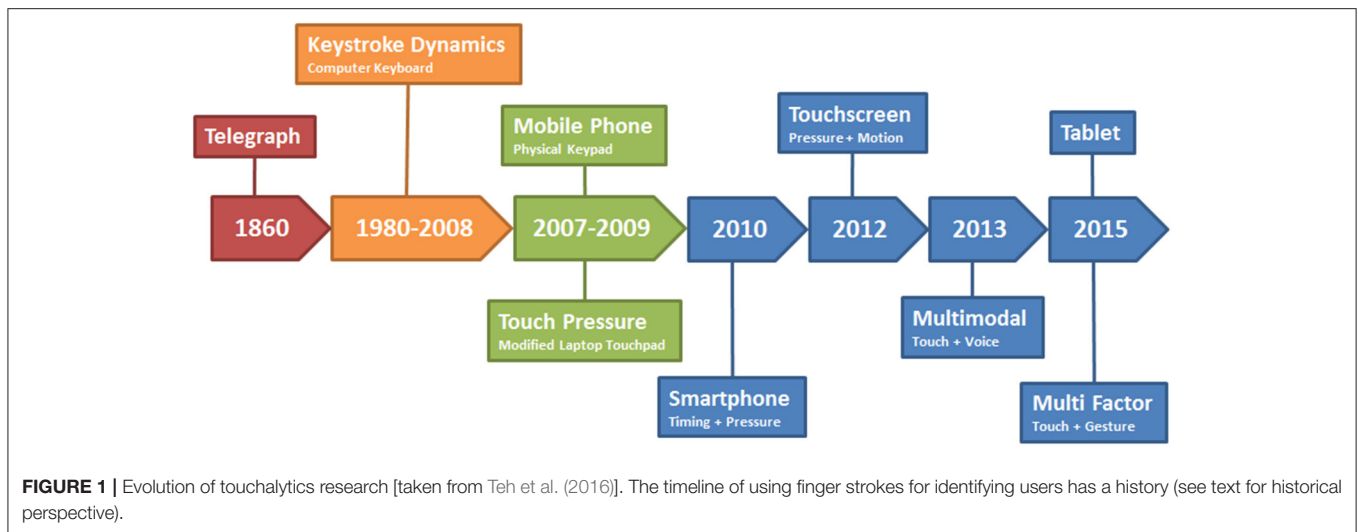
erect telephone poles along the battle routes (Jenkins et al., 2011). Therefore, field units used Morse Code over portable two-way radios to communicate with each other. However, the risk of enemy eavesdropping and intercepting the wireless communications persisted and radio operators had to remain anonymous. Consequently, the army further developed the behavior biometric of identifying a radio operator by the gaps in their taps on the radio (Jenkins et al., 2011).

The 1970's saw the rise of computer terminals with operators sitting at keyboards and entering commands and typing messages. Consequently, IBM researchers considered that keyboard typing rhythms can be used to distinguish typists. Spillane asserted this in a bulletin, but did not mention any experiments which were conducted (Spillane, 1975; Killourhy, 2012). However, Forsen et al. (1977) conducted several tests in order to determine whether or not the typing patterns of one typist can be determined. Among the experiments conducted was one where a small group of subjects were asked to type their own and each other's names. The researchers presented findings indicating that it was possible to identify the subjects typing their own names from those who were not (Killourhy, 2012).

A more in-depth study was conducted by Gaines (Gaines et al., 1980) using seven typists to transcribe three passages of words and sentences while monitoring the time between when a key was pressed and when it was released. From this timing, they were able to analytically determine differences among typing patterns. The researchers found that time between two consecutive keystrokes follow a log-normal distribution. Moreover, they were able to develop a statistical test which was able to successfully distinguish which transcript was typed by each of the seven typists. However, the authors acknowledged that more research would be needed in order for their findings to be conclusive, since their experiment was conducted on a small population (Gaines et al., 1980; Killourhy, 2012). The 1990's and early 2000's saw researchers endeavoring to use machine learning algorithms to classify whether or not a series of keystrokes belonged to a particular user. Brown and Rogers (1994) utilized a neural network to identify imposters with a 0% miss rate and a 12% false positive rate. Moreover, Azevedo et al. (2007a,b) modified a Support Vector Machine classifier which utilized genetic algorithms and particle swarm optimization and thus achieved miss and false positive rates between 1.1 and 1.2% (Killourhy, 2012).

The touch screen was developed in the 1960's for air traffic control systems (Ion, 2013). Today, they are found in ATM machines, self-service kiosks, and most notably, the smartphone. Moreover, the internet has replaced the letter, telephone, telegraph, and desktop computers for communication. Today, the smartphone is the most commonly used communication device and is always connected to the internet. People store a variety of personal and sensitive information on their smartphones from credit card information, photographs, fingerprints, emails, and text messages. Hence, this development has provided the impetus for adversaries to conduct their activities on a much wider scale. In fact, one in three mobile phone users has experienced device theft (Norton, 2011). Gaining access to a person's mobile phone can allow a criminal





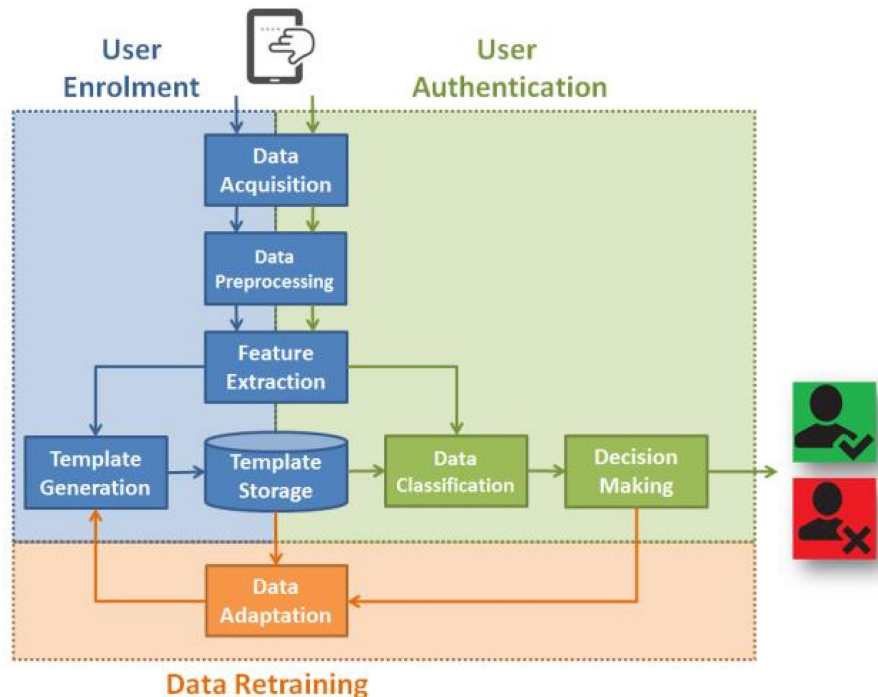
the opportunity to commit a variety of dangerous actions such as impersonating the legitimate user and stealing that person's identity.

When a Personal Identification Number (P.I.N) is used for identification, an intruder can possibly steal the PIN by looking over the shoulder of a user logging into a system, i.e., “shoulder surfing.” As a result of the greater use of the touchscreen, biometric verification methods for mobile devices are becoming an important tool to ensure security for usage of the mobile devices (Trewin et al., 2012). Conventional sensors are fingerprint sensors, retina scanners, and facial recognition cameras, which are now embedded in a large array of devices (Trewin et al., 2012). However, many of these biometric verification methods are single use authentication schemes, which can be compromised by other users and then be used to access sensitive information. Fingerprint identification can be compromised by activating the device while a person is asleep. For example, a suspicious wife unlocked her sleeping husband's phone on an airplane and found information implicating her spouse in an extramarital affair (Sky-News, 2017). Even retinal and facial recognition systems can be compromised by malicious or coerced intruders. In fact, hackers were able to trick the retinal scanner on Samsung's flagship Galaxy S8 phone within 1 month of the phone's release (Hern, 2017). Furthermore, attackers are becoming more adept at using social engineering to elicit passwords and other authentication means from unsuspecting individuals such as seniors and teenagers. That is, attackers utilize spear-phishing and are able to convince users to download malware onto computers. This malware then grants the attackers access to the victims' machine (Ariu et al., 2017). In general, once the device is compromised, it would allow the intruder unfettered access to the phone's contents, which could have serious consequences.

Prior to the proliferation of the touchscreen smartphone, mobile phones with physical keypads were the norm (McCloughlin and Naransamy, 2009). These devices were secured by P.I.N

codes which the user had to physically type on the keypads. However, the keypad can only record the timing interval of the keystrokes and thus this severely limits the ability of a classifier to authenticate a user using this biometric feature (Teh et al., 2016). In fact, while mobile keypad analytics may result in lower accuracies than that of other biometric methods, it has been asserted by Ali et al. (2017) that such an approach is nevertheless economic, noninvasive, and provides an opportunity for continuous authentication.

Recently, there has been a focus on co-adaptation and Embodiment as a means for human interaction with robotic systems (Beckerle et al., 2019). Co-adaptation is where the machine adapts to the human operator, while the human adapts to it. Embodiment is where the participants are equipped with the investigated technical device and become part of the human-machine system and its control loop. Due to the direct interaction of both partners, such experiments can yield detailed information during the usage, which might not be captured by post experiment questionnaires or measures. This might even be performed during development to guide the design in a user-centered direction (Beckerle et al., 2019). It has been suggested that Co-adaptation and Embodiment can have a more general framework (Ziemke, 2003). One way to look at embodiment is as a structural coupling (Ziemke, 2003). A system “X” is embodied in an environment “E” if perturbatory channels exist between the two. That means, “X” is embodied in “E” if for every time “t” at which both “X” and “E” exist, some subset of E's possible states with respect to “X” have the capacity to perturb X's state, and some subset of X's possible states with respect to “E” have the capacity to perturb E's state (Ziemke, 2003). From this perspective biometric security methods in user interfaces would be important for cybersecurity. Co-adaptation can be imagined as a transformation of a “simple tool” to that of an “intelligent tool.” As a result of this transformation, the new tool becomes pivotal to the defining and accomplishing the user's security goal (Sanchez et al., 2009). An application of these ideas can be used



**FIGURE 2 |** A touchalytics framework for identifying a user [taken from Teh et al. (2016)]. (1) The data is acquired from finger swiping of a mobile device. (2) The data is pre-processed. (3) Features are extracted. (4) The data are classified. (5) A decision about whether the user is appropriate or not. The decision is fed back and the template for the user is adapted and stored. Taken from Teh et al. (2016).

to build better security systems for user interaction with smart devices. The user's goal in communicating with a smart device is to maintain security.

Saevanee and Bhatarakosol (2008) simulated a touch screen by collecting finger stroke data from a laptop's touchpad. Users were asked to enter various 10-digit codes using the touch pad entirely. The features extracted were the inter stroke time, the hold time (time which a finger was on the keypad) and the finger pressure. Finger pressure, according to Teh et al. (2016) can be determined by the device's operating system. A user's touch pressure value is directly related to the strength of that user's finger muscle. Therefore, the touch pressure is unique to each user and is very difficult for a shoulder surfing adversary to imitate by mere observation (Teh et al., 2016). In fact, Saevanee and Bhatarakosol (2008) reported that when finger pressure was used as the input to a KNN classifier, 99% accuracy was reported. Moreover, when all three of the aforementioned features were considered the accuracy fell to 90%. Frank et al. (2013) endeavored to understand how many finger swipes/strokes contributed to better performance rates. They postulated that a series of strokes belonging to an individual user resulted in better results than that of a single stroke. Their experiments yielded a 13% equal error rate (EER) for a single stroke and 2–3% for 11–20 (inclusive) strokes (Frank et al., 2013). Miguel-Hurtado et al. (2016) presented findings which indicated that it was possible to predict the sex of a mobile user based on finger strokes. Their experiment utilized the SSD dataset

which consisted of 116 users (Guest et al., 2014). By utilizing Naïve Bayes and Logistic classifiers, the researchers were able to make predictions with ~78% accuracy. Wang et al. (2017) endeavored to utilize the same biometric method to augment continuous authentication and cross device authentication by conducting a similar study. That is, the researchers proposed transferring a behavioral model from one device to others in order to compare results. Subjects were tasked with interacting with a News application on different devices and various touch stroke characteristics were collected including the X, Y coordinates of their fingers on the screen, timestamp, pressure, and finger size. Consequently, an area under curve score of 80–96% was achieved by utilizing SVM and Random Forest classifiers (Wang et al., 2017).

One way to have an extra level of security and protect the system is to implement a continuous authentication method, without imposing a burdensome requirement of having to re-enter the identification. One such methodology has been developed and is known as Touch analytics (Frank et al., 2013). This technique is the process of user authentication based on finger movements on a touchscreen. Each user has a unique way in which he uses a mobile phone's touchscreen. For example, the way one user's fingers swipe a touch screen is different from that of another (Frank et al., 2013) and could be the basis of devising continuous monitoring of user-phone interaction and making it more secure. The swipe is one of the most frequent methods in which a user interacts

Publication	Year	Subject Size	Input Device Information				Performance		
			User Input	UC	Device Used	Display Size	Features	Classification	EER(%)
Frank et al. [78]	2013	41	A	-	-	-	X, Y, timestamp, orientation	SVM RBF Kernel and k-NN	13 single stroke 2-3, 11 to 12 strokes up to 20
Li et al. [84]	2013	75	A	-	Motorola Android phones	480 × 854 pixels	X, Y, pressure, distance, time	SVM Gaussian Radial Basis function	Portrait slide up-95.78%, Landscape slide down- 94.20%
Xu et al. [4]	2014	30	A	-	Galaxy SII	4.3	X, Y, timestamp, size, pressure	SVM RBF	30 users- pinch (3.33%) slide (1.3%)
Feng et al. [17]	2014	23 phone owners and 100 guest users	P	U	8 Samsung Galaxy S-III, 3 Galaxy S-IV and 12 Nexus 4 S3	4.8, SIV, Nexus IV- 4.7	X, Y, timestamp, size, pressure, swipe length, swipe curvature	DTW with One Nearest Neighbor	90% accuracy
Zheng et al. [3]	2014	80	A	C	Samsung Galaxy Nexus	4.65	Acceleration, pressure, size, and time	Nearest Neighbor distance	3.65
Bo et al. [2]	2014	100	P	C	HTC EVO 3D and Samsung Galaxy S3	HTC -4.3, S3-4.8	X,Y, timestamp, pressure, vibration, rotation - static and motion modes	SVM	Static scenario (FAR Tap 22, Fling 9, Scroll-23), Walking scenario Accuracy 100% after 12 steps of walking
Zhao et al. [27]	2014	78	A	C	Samsung Galaxy S3	4.8	X, Y, Pressure, timestamp	STDI with GTGF	
Saravanan et al. [83]	2014	20	A	C	Nexus 7, Nexus 4	Nexus 4 - 4.7, Nexus 7 - 7	X, Y, Pressure, relative timestamp	SUA -SVM and RF MUA NB, J48, Random Forests and BayesNet	97.9% accuracy mobile phones 96.79% - Tablets
Zhang et al. [29]	2015	50	A	C	iPhone 5s	4	X, Y, timestamp	SRC, rbfsVM & KSRC	rbfsVM - 19 swipes - 1.13 0.22
Miguel-Hurtado et al. [81]	2016	116	A	C	Galaxy S2	4.3	X, Y, timestamp, pressure, size	NB, logistic regression, SVM and decision tree	78 % accuracy rates
Sharma et al. [85]	2017	42	A	C	Google Nexus 7	7	X, Y, timestamp, pressure, size	SVM	Two class SVM - 7
Ahmad et al. [86]	2017	40	A	C	-	-	Interaction trace map	SVM	All interactions - 80.27
Wang et al. [77]	2017	160 set of app usage data	P	U	Nexus S, Nexus 4, Nexus 7-2012, Nexus 7-2013	Nexus S-4,Nexus 4-4.7, 7- 7	X, Y, timestamp, pressure, size	SVM & RF	AUC score of 80% to 96% (detecting unauthorised access)

**FIGURE 3 |** The progression of touch dynamics research, which has evolved into what is presently called touchalytics. Taken from Ellavarason et al. (2020).

with a smartphone. In fact, research has proven the swipe to be a reliable means of identifying users (Ellavarason et al., 2020). In real time, the swiping can adapt to the touchscreen while the touchscreen's intelligent interface adapts to the human swiping, implementing co-adaptation (Beckerle et al., 2019).

## The Intricacies of a Touch Dynamics System

According to Teh et al. (2016), the schematic of any touch dynamics system follows the diagram outlined in **Figure 2**. That is, before any type of authentication can be done, there must be data. In any smartphone, the touch-strokes are recorded using

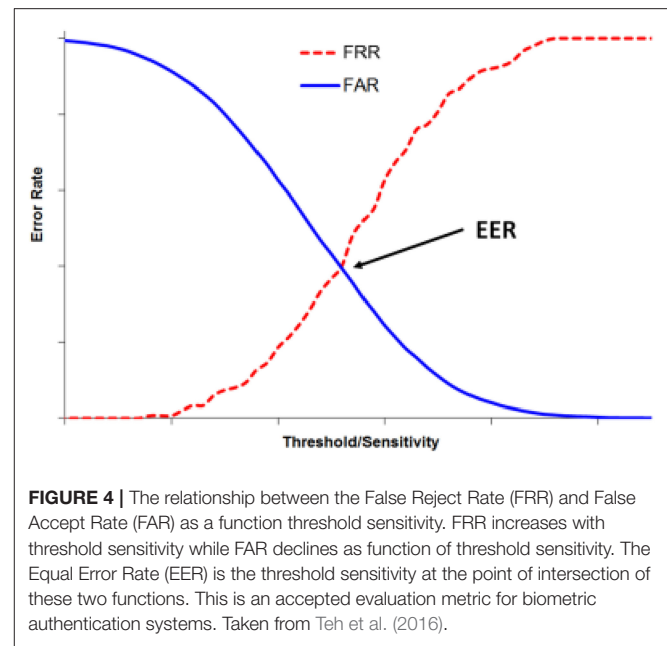
sensors embedded in the touchscreen. Next, the data needs to be preprocessed or cleaned in order to remove error values and outliers which can mislead the classifier during training. Thereafter, feature extraction is done in order to reduce the effects of dimensionality (Teh et al., 2016). Accordingly, the model is then trained on the training data and then evaluated on the test set. However, given the continuous nature of the proposed authentication mechanism on a smartphone, there is an error correction loop in the process. That is, after the model makes a decision, it continues to collect data in the form of touch strokes and attempts to retain itself and adapt using the new data (Teh et al., 2016).

## Motivations

The objective of our study is to investigate how the utilization of a broad range and continuous monitoring of user interaction with a device can be made fast enough to ensure that the owner of the device is using it and the control of the device has not been compromised. However, the number of features defined are relatively large and could compromise the performance of the system because the KNN algorithm used by Frank et al. (2013) suffer from the curse of dimensionality (Yiu, 2019).

All the major studies conducted on touch stroke dynamics on mobile devices was outlined (Ellavarason et al., 2020; **Figure 3**). From this table (**Figure 3**), we can see that the primary classification algorithms used are Support Vector Machines, Nearest Neighbors, Decision Trees Random Forests, Logistic Regression, and Bayesian models. Touch strokes on a mobile device often happen in a sequence of actions. That is, almost everyone utilizes more than one touch stroke when interacting with a smartphone. Therefore, the nature of this process makes it a prime candidate for time series analysis. For example, when given “n” consecutive touch strokes of a user, what is the probability that the “n+1” stroke will belong to that said user? From our review of the literature, we found no time series approach taken to solve the authentication problem outlined by Frank et al. (2013) and is a possible extension from that proposed in this paper.

The purpose of this paper is to explore the present effectiveness of Touch-analytics and consider how this might be improved using Principal Component Analysis (PCA) and compare the performance of the reduced feature space to that of having all the features. PCA analysis computes the eigenvectors of the covariance matrix of the defined touchanalytics features. This covariance matrix is symmetric and generates an orthogonal basis of PCA vectors, which can be used to optimize the important features. There are algorithms that can compute PCA in real time using Oja's rule in a neural net implementation of Hebbian learning (Oja, 1989). A robust continuous authentication system will give users an extra layer of security and an increased sense of peace of mind if their devices are lost or stolen. Consequently, such improvements using PCA may prevent access to sensitive information and thus will save businesses money (Lau, 2018). If continuous authentication models become successful and easily



implementable, it would inhibit the growing problem of mobile device theft (Norton, 2011).

## Methodological Approach

One particular measure of the effectiveness of the authentication procedure is to use Decision Theory to reduce the false acceptance vs the false rejection of a user (Powers, 2020). One such metric has been defined as the Equal Error Rate (EER) as the intersection point of the false rejection and false acceptance (Kar-Ann, 2008; **Figure 4**). In this paper, we develop and implement a touch classifier that improves the EER of a single stroke mobile authentication system. The research described in this paper lends itself to further enhancement using neural network authentication schemes.

The Frank et al. (2013) study implemented two classifiers in order to determine if a series of stroke patterns belonged to a particular user. These classifiers implemented the K-Nearest Neighbors and the Support Vector Machines algorithm. Frank et al. (2013) to identify a particular user.

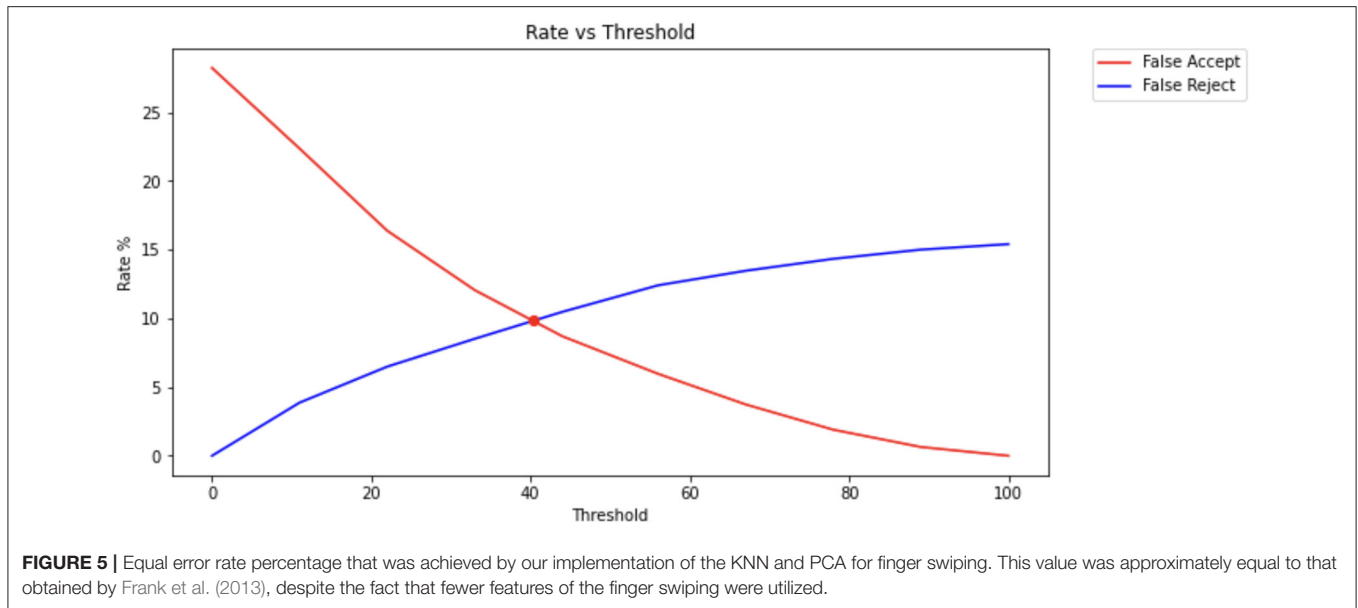
## RESULTS

The model of Frank et al. (2013) has postulated that there are 30 features, which can be derived from a single stroke by a finger on a touch screen (see **Appendix B** for this list). As a first pass, we implemented the KNN and PCA algorithms as a basis for comparison to those of Frank et al. (2013). A python implementation is given in **Appendix A**.

## KNN Implementation

We first tested the efficacy of all the features by implementing a K-Nearest Neighbors algorithm and compared it with the results obtained by Frank et al. (2013) who used MATLAB in their





implementation. This was used as a basis for comparing their results with our implementation of this system and finding the PCA components (**Figure 5**). We utilized the KNN clustering algorithm because Frank et al. (2013), used it in their studies and we wished to insure that the dimensionality reduction that we found from the PCA analysis of the finger swiping data was not affected by the clustering algorithm used. This was important for comparison purposes.

We implemented a principal component (PCA) program (**Appendix A**) to extract the dominant features from the given 30 features and compared the results to those using the K-nearest neighbor algorithm.

## PCA-Based Algorithm Design and Implementation

It is possible to determine which features in a dataset are more relevant when performing principal component analysis process. Hence, the phone ID, User ID, and document ID fields were removed from the dataset as (Frank et al., 2013) indicated that such were not to be used for testing. Feature Scaling was then performed on the data using the standardization technique. A covariance matrix was then extracted from the dataset using the scikit-learn library in Python (Pedregosa et al., 2011). This was followed by an eigenvector decomposition from this matrix using the scikit-learn Python library function.

The distribution of the variance among all the Principal Components derived from the dataset with descending variance shows that the first 15 principal components contain most of the variance (**Figures 6, 7**). Among these principal components, we consistently determined the following dominant features that can be used in computing the equal error rate:

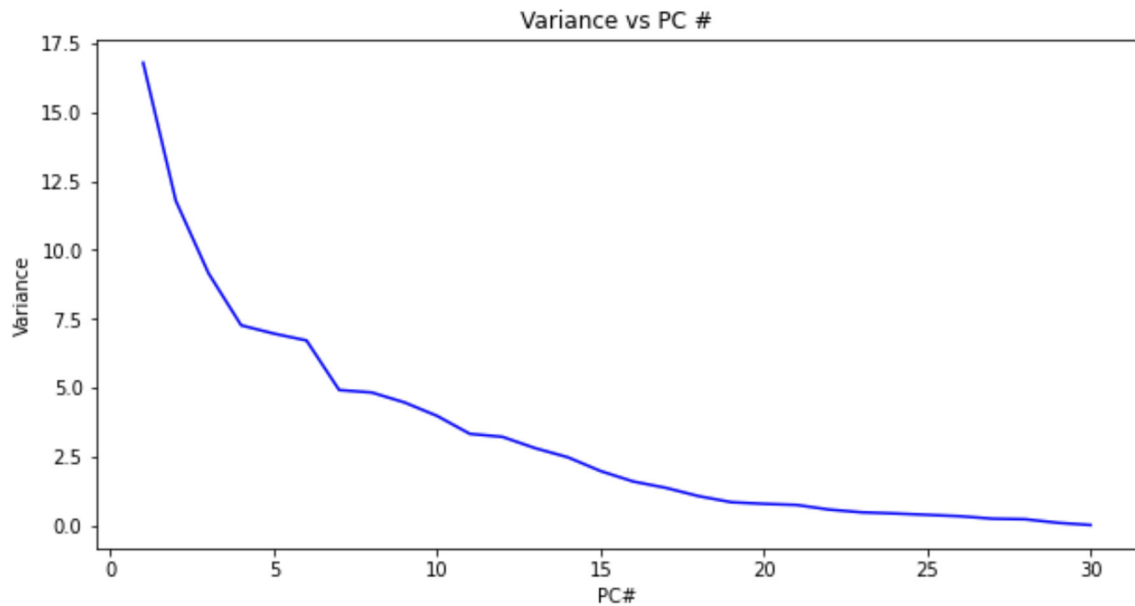
1. average velocity
2. largest deviation from end-to-end line
3. 50%-perc. pairwise acc

4. length of trajectory
5. start \$y\$
6. mean resultant length
7. up/down/left/right flag
8. start \$x\$
9. mid-stroke area covered
10. inter-stroke time
11. mid-stroke pressure
12. mid-stroke finger orientation
13. stroke duration
14. 20%-perc. pairwise velocity
15. phone orientation.

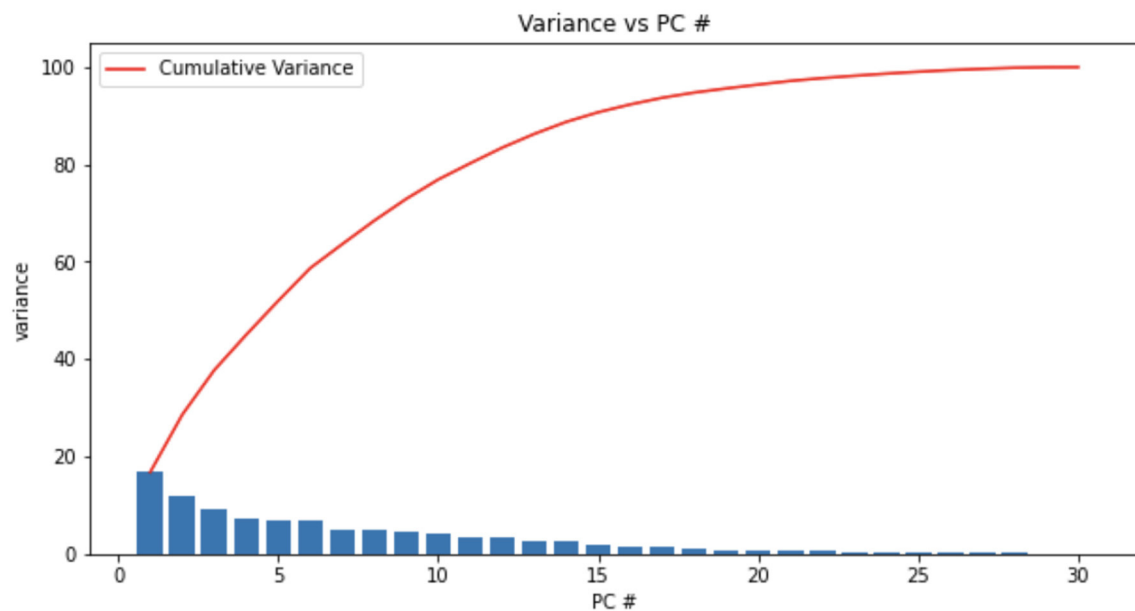
Therefore, we can conclude that these 15 features are most helpful in uniquely identifying a user on a mobile device. The full rankings of the features up to 23 can be found in **Appendix C**.

## Performing Cross-Validation on Touchalytics Dataset

The authentication problem which we are trying to solve can be described as a two class classifier (Bishop, 2006). That is, the classifier should determine whether the user is authorized to use the device or is not authorized to use the device using features representing the touch patterns of 40 different users interacting with five different documents on a mobile device. We start by selecting all of the strokes belonging to a particular user and recording a count of such. These strokes are given a label of "Class One." We then randomly select the same number of strokes belonging to the other users. These are given the label "Class Two" Hence, we have the same number of strokes belonging to Class One and Class Two. The data was then cleaned by removing the rows with missing and infinite values. In addition, the class column was designated the target variable "Y" and was removed. Furthermore, the doc id, user id, phone id, and class were removed from the data set as these were deemed to not have



**FIGURE 6** | Variance as a function of principal component  $f(PC\#)$  using the data from Frank et al. (2013). The variance orders the principal components. The cumulative variance reaches a plateau and is a measure of the number of PC's necessary to contain the information necessary to identify a pattern of finger swiping.



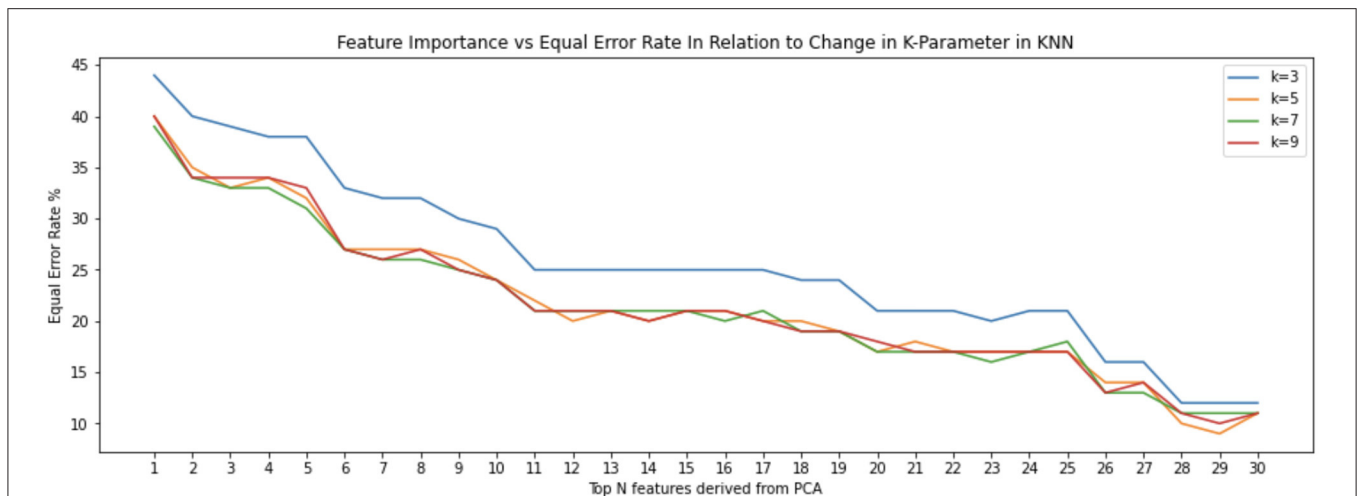
**FIGURE 7** | The variance vs. the 30 principal components. The cumulative variance reached a plateau at close to 15 principal components. We found that 23 of the 30 principal components were the most prominent. See text for details.

any impact on the results by Frank et al. (2013). The remaining columns were designated as the variable X.

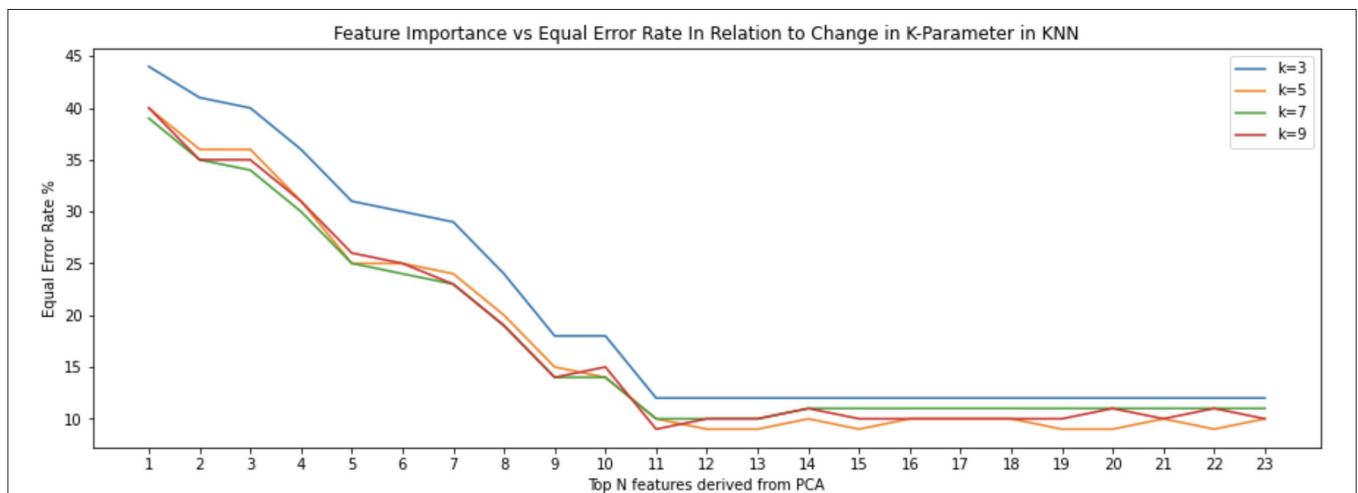
Accordingly, we then split the dataset into training (80%) and testing (20%) sets. Consequently, the data were rescaled using the MiniMax Method. A KNN model was created utilizing a library from Scikit learn (Pedregosa et al., 2011). We chose 3, 5, 7, and 9 neighbors respectively as the “K” parameter and trained our model. Thereafter, we made predictions on the test data.

## DISCUSSION

This study has shown that the main feature for characterizing touch dynamics for touchscreen mobile devices (Frank et al., 2013) is the average velocity of a finger stroke. The novelty in our research is in our comparison of the effectiveness of the different features toward the creation of a digital fingerprint of a mobile phone user. Using the database established by Frank et al. (2013),



**FIGURE 8 |** Classification done using all 30 features defined by Frank et al. (2013), derived from the first principal component result in stabilization of equal error rate. These features are numbered according to their order of importance: “average velocity,” “50%-perc. pairwise velocity,” “median velocity at last 3 pts,” “80%-perc. pairwise velocity,” “20%-perc. pairwise velocity,” “direction of end-to-end line,” “average direction,” “80%-perc. pairwise acc,” “stroke duration,” “direct end-to-end distance,” “stop \$y\$,” “20%-perc. pairwise acc,” “ratio end-to-end dist and length of trajectory,” “median acceleration at first 5 points,” “50%-perc. pairwise acc,” “length of trajectory,” “largest deviation from end-to-end line,” “phone orientation,” “80%-perc. dev. from end-to-end line,” “stop \$x\$,” “50%-perc. dev. from end-to-end line,” “20%-perc. dev. from end-to-end line,” “start \$y\$,” “mid-stroke finger orientation,” “up/down/left/right flag,” “mid-stroke area covered,” “mean resultant length,” “mid-stroke pressure,” “start \$x\$,” and “inter-stroke time.”

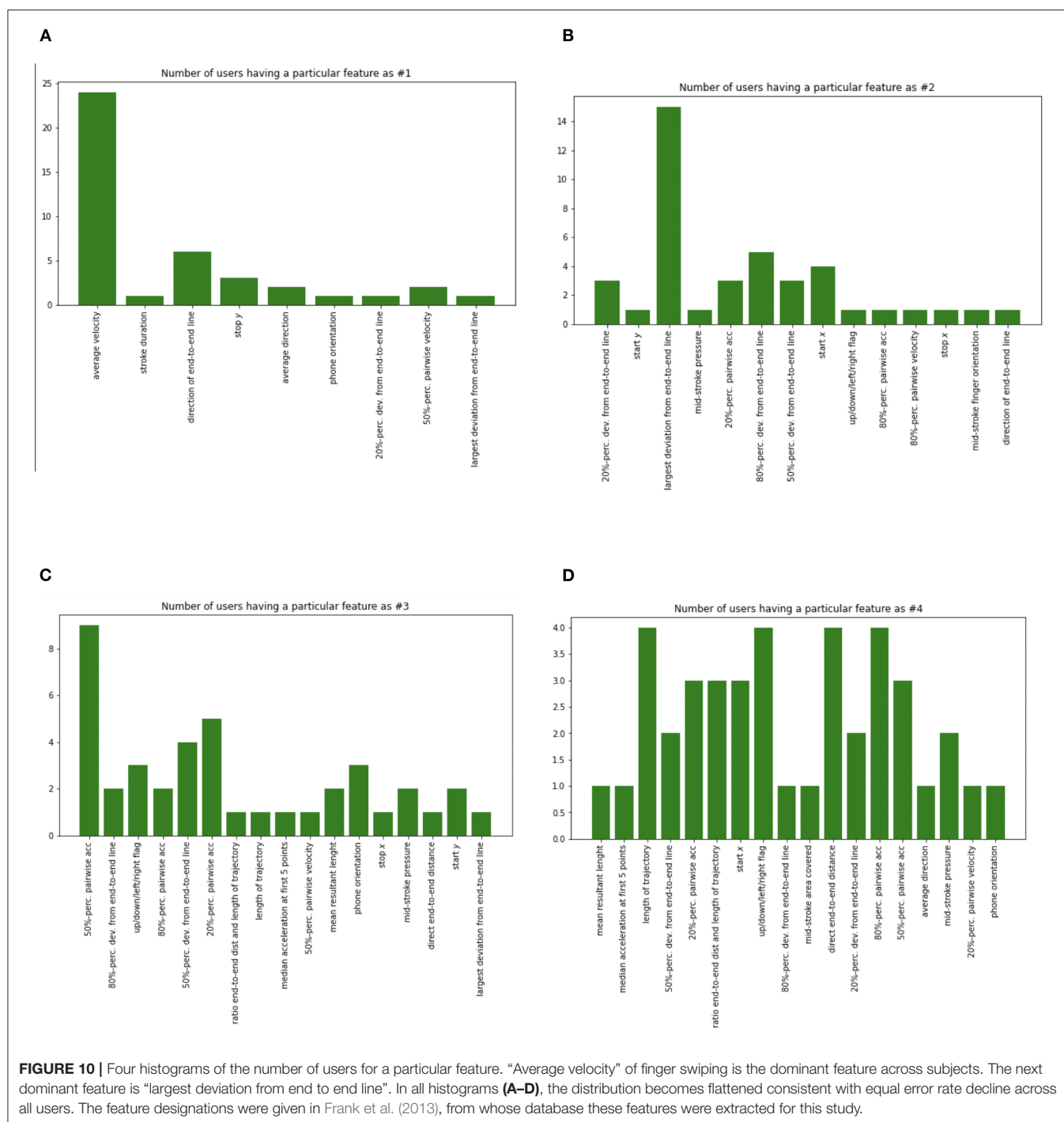


**FIGURE 9 |** Classification done using the 23 features derived from the top 30 principal components result in a more pronounced stabilization of equal error rate. These features are numbered according to their order of importance are: “average velocity,” “largest deviation from end-to-end line,” “50%-perc. pairwise acc,” “length of trajectory,” “start \$y\$,” “mean resultant length,” “up/down/left/right flag,” “start \$x\$,” “mid-stroke area covered,” “inter-stroke time,” “mid-stroke pressure,” “mid-stroke finger orientation,” and “stroke.”

we have expanded on this to suggest that by using co-adaptation of user and finger swiping, we can incorporate continuous authentication to enhance security. Similar behavioral biometric features were used to characterized telegraph communication (Jenkins et al., 2011), keyboard typing (Gaines et al., 1980), and keypad typing (Teh et al., 2016) using ad hoc approaches. That is, the speed at which the user's fingers move on the touch screen is what is most unique to each user. This is most consistent with the “Behavioral Biometrics” theme of this paper, which is

based on the idea that each human being has a unique behavioral identifiable characteristic when typing (Horst et al., 2019). As a result, this feature would be integral in the building of any finger stroke touch dynamics classification model.

History supports this observation. While the devices have changed over time, the biological composition of a human being has remained the same. That is, in most devices which requires the user to utilize their fingers, the speed in which a human press and releases a surface remained unique to that user. This



is a result of the unique muscle composition, nerve organization and electrical impulses connections among the brain, nerves, and muscle connections that are learned over time (Bhattacharyya et al., 2009). During the days of Morse Code and the telegraph, operators were able to identify their counterparts by the rate of the taps on the medium. Approximately, half a century later, Gaines et al. (1980) observed that the intervals between a typist's keystrokes could be modeled using a log normal distribution. More recently, Teh et al. (2016) reported that they could identify

users based on the intervals of key presses on the physical keypad of a mobile device. Furthermore, and most relevant to our paper, a plethora of research conducted on touch patterns on mobile devices have come to a similar conclusion.

With the advent of touch screen devices, the user now has to move his finger over a larger surface area. Consequently, the start and end positions of the user's finger strokes have become a set of new features which can possibly be used to identify an individual. However, it must be noted that the end positions of a finger stroke



were not paramount in the top principal components. Therefore, we can assert that the end position of a finger stroke should not be used as the sole basis for identifying a user. However, when these features are combined with the aforementioned velocity measures, the results could be more beneficial and less prone to intrusion.

We then sorted the principal components according to their largest eigenvalue (variance) (**Figure 7**). We further sorted the 30 features (Frank et al., 2013) from the strongest feature to the weakest in the first principal component. This showed a steady decline in EER as we added features for each  $k$  in the KNN algorithm all the way up to 30 features (**Figure 8**). KNN was utilized as the clustering algorithm, since that was the one used by Frank et al. (2013), so that there would be an adequate basis for comparison.

However, by using additional principal components, we discovered that there was a plateau such that strongest 12 features resulted in similar or better results as when all 30 are used with a decline in EER (**Figure 9**). There are repeats of seven of the 30 most important features. Therefore, 23 features are the most relevant in the top 30 principal components. This is significant because it means that an authentication system can remain effective using fewer features. That is, if the system has to consider fewer features when performing a classification, less sensors have to be used and thus energy is saved. In addition, if the classifier has to consider less data, then it becomes more efficient. What is surprising is the fact the  $K$ -value chosen for the KNN classifier doesn't seem to affect the Equal Error Rate. This is depicted in the interwoven KNN curves (**Figures 8, 9**). Furthermore, we can observe that the equal error rate continually drops after each one of the top 15 features is considered and then remains within a certain range or reaching a plateau (**Figure 9**).

We also plotted a histogram of the number of users for a particular feature (**Figure 10**). The analysis indicates that "average velocity" of finger swiping is the dominant feature across subjects. The next dominant feature is the "largest deviation from end to end line." Four of these histograms (**Figures 10A–D**) show that as we consider more features, the distribution becomes flattened, consistent with the equal error rate decline across all users. The equal error rate decline and stabilization indicate that using a 15 dimensional feature vector well characterizes each user.

It would be of interest to compare datasets from other touch screens such as kiosks etc., which have larger screens in order to learn if the start and stop positions are consistent

with that of users on a personal mobile device. Similar to how the inter-stroke time and stroke duration were consistently used to discern users among all touch-based devices in the past, we expect that the stop and start positions would also be consistent among all touch screen devices. Thus, the application of the principles of embodiment and co-adaptation could be important in the development of secure and efficient ways of human interaction with mobile communication devices.

## DATA AVAILABILITY STATEMENT

The data was obtained from the dataset created by Frank et al. (2013).

## AUTHOR CONTRIBUTIONS

AJ contributed to the conceptual understanding of the biometric features, writing the computer programs to implement the algorithms used in the study, and writing of the manuscript. TH contributed to some of the concepts utilizing touchalytics, reviewed some of the code for the implementation of the touchalytics algorithm, and contributed to the writing of the manuscript. TR contributed to the organization, writing of the manuscript, and utilization of machine learning approach to finding the dominant bio-metric features of the classification algorithm. All authors contributed to the article and approved the submitted version.

## FUNDING

TR has been funded by City University of New York through a Distinguished Professorship, which allows for re-assigned time to allow him to do this research. AJ has received support from a grant to Brooklyn College from New York State to support the Star Early College program (TR) and Research funds that support the research of TR. TR was also funded by an award made by Pheobe Cohen to Dr. Raphan's General Research Activities-7B613-00-01.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomp.2022.754716/full#supplementary-material>

## REFERENCES

- Ali, M. L., Monaco, J. V., Tappert, C. C., and Qiu, M. (2017). Keystroke biometric systems for user authentication. *J. Signal Process. Syst.* 86, 175–190. doi: 10.1007/s11265-016-1114-9
- Ariu, D., Frumento, E., and Fumera, G. (2017). "Social engineering 2.0: a foundational work: invited paper," in *Proceedings of the Computing Frontiers Conference*. (Siena: Association for Computing Machinery). doi: 10.1145/3075564.3076260
- Azevedo, G., Cavalcanti, G., and Filho, E. (2007a). "An approach to feature selection for keystroke dynamics systems based on PSO and feature weighting," in *In IEEE Congress on Evolutionary Computation (CEC 2007)*. Piscataway, NJ: IEEE. doi: 10.1109/CEC.2007.4424936
- Azevedo, G., Cavalcanti, G., and Filho, E. (2007b). "Hybrid solutions for the feature selection in personal identification problems through keystroke dynamics," in *International Joint Conference on Neural Networks (IJCNN 2007)*. Orlando, FL: IEEE. doi: 10.1109/IJCNN.2007.4371256
- Beckerle, P., Castellini, C., and Lenggenhager, B. (2019). Robotic interfaces for cognitive psychology and embodiment research: a research roadmap. *WIREs Cogn Sci.* 10, e1486. doi: 10.1002/wcs.1486
- Bhattacharyya, D., Ranjan, R., Alisherov, F., and Minkyu, C. (2009). Biometric authentication: a review. *Int. J. u- e- Service Sci. Technol.* 2, 82–86.

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin: Springer-Verlag.
- Brown, M., and Rogers, S. J. (1994). "A practical approach to user authentication," in *10th Annual Computer Security Applications Conference*. Orlando, FL: IEEE Computer Society.
- Cho, C., Kunin, M., Osaki, Y., Olanow, C. W., Cohen, B., and Raphan, T. (2006). "A model-based approach for assessing Parkinsonian gait and effects of Levodopa and Deep Brain Stimulation," in *Proc. 28th IEEE EMBS Annual International Conference*. New York City, NY. doi: 10.1109/IEMBS.2006.259439
- Cho, C., Kunin, M., Osaki, Y., Olanow, C. W., Cohen, B., and Raphan, T. (2010). Frequency-velocity mismatch: a fundamental abnormality in Parkinsonian gait. *J. Neurophysiol.* 103, 1478–1489. doi: 10.1152/jn.00664.2009
- Dooley, J. (2013). *A Brief History of Cryptology and Cryptographic Algorithms*. Springer doi: 10.1007/978-3-319-01628-3
- Ellavarason, E., Guest, R., Deravi, F., Sanchez-Riello, R., and Corsetti, B. (2020). Touch-dynamics based behavioural biometrics on mobile devices – a review from a usability and performance perspective. *ACM Comput. Survey (Rome)*, 53:120. doi: 10.1145/3394713
- Forsen, G., Nelson, M., and Staron, R. J. (1977). *Personal Attributes Authentication Techniques*. (Rome, NY: Pattern Analysis & Recognition Corp).
- Frank, M., Biedert, R., Ma, E., Martinovic, I., and Song, D. (2013). Touchalytics: on the applicability of touchscreen input as a behavioral biometric for continuous authentication. *IEEE Trans. Inform. Forens. Security* 8, 136–148. doi: 10.1109/TIFS.2012.2225048
- Gaines, R., Lisowski, W., Press, S., and Shapiro, N. (1980). *Authentication by Keystroke Timing: Some Preliminary Results*. Santa Monica, CA: RAND Corporation.
- Guest, R., Miguel-Hurtado, O., Stevenage, S. V., Neil, G. J., and Black, S. (2014). "Biometrics within the SuperIdentity project: a new approach to spanning multiple identity domains," in *2014 International Carnahan Conference on Security Technology (ICCTST)* (Rome), 1–6. doi: 10.1109/CCST.2014.6986992
- Hern, A. (2017). *Samsung Galaxy S8 Iris Scanner Fooled by German Hackers*. Available online at: <https://www.theguardian.com/technology/2017/may/23/samsung-galaxy-s8-iris-scanner-german-hackers-biometric-security> (accessed January 5, 2021).
- Horst, F., Lapuschkin, S., Samek, W., Müller, K., and Schöhlhorn, W. (2019). Explaining the unique nature of individual gait patterns with deep learning. *Sci. Rep.* 9, 8. doi: 10.1038/s41598-019-38748-8
- Ion, F. (2013). *From Touch Displays to the Surface: A Brief History of Touchscreen Technology*. arsTechnica. Available online at: <https://arstechnica.com/gadgets/2013/04/from-touch-displays-to-the-surface-a-brief-history-of-touchscreen-technology/> (accessed July 20, 2021).
- Jenkins, J., Nguyen, Q., Reynolds, J., Horner, W., and Szu, H. (2011). "The physiology of keystroke dynamics," in *Proceedings of SPIE - The International Society for Optical Engineering*, 8058. doi: 10.1117/12.887419
- Kar-Ann, T. (2008). "Between AUC based and error rate based learning," in *2008 3rd IEEE Conference on Industrial Electronics and Applications*, 2116–2120. doi: 10.1109/ICIEA.2008.4582893
- Killourhy, K. S. (2012). *A Scientific Understanding of Keystroke Dynamics*. Pittsburgh, PA: Carnegie Mellon University.
- Lau, L. (2018). *Cybercrime "Pandemic" May Have Cost the World \$600 Billion Last Year*. CNBC. Available online at: <https://www.cnbc.com/2018/02/22/cybercrime-pandemic-may-have-cost-the-world-600-billion-last-year.html> (accessed September 30, 2020).
- Ma, Y., Huang, Z., Wang, X., and Huang, K. (2020). An overview of multimodal biometrics using the face and ear. *Hindawi Math. Probl. Eng.* 2010, 1–17. doi: 10.1155/2020/6802905
- McLoughlin, I. V., and Naransamy, M. S. (2009). "Keypress biometrics for user validation in mobile consumer devices," in *2009 IEEE 13th International Symposium on Consumer Electronics (Kyoto)*, 280–284. doi: 10.1109/ISCE.2009.5156933
- Miguel-Hurtado, O., Stevenage, S. V., Bevan, C., and Guest, R. (2016). Predicting sex as a soft-biometrics from device interaction swipe gestures. *Pattern Recogn. Lett.* 79, 44–51. doi: 10.1016/j.patrec.2016.04.024
- Norton (2011). *Norton Survey Reveals One in Three Experience Cell Phone Loss, Theft*. Norton. Available online at: [http://www.symantec.com/about/news/release/article.jsp?prid=20110208\\_01](http://www.symantec.com/about/news/release/article.jsp?prid=20110208_01) (accessed January 5, 2021).
- Oja, E. (1989). Neural networks, principal components, and subspaces. *Int J Neural Syst.* 1, 61–68. doi: 10.1142/S0129065789000475
- Olanrewaju, L., Oyebiyi, O., Misra, S., Maskeliunas, R., and Damasevicius, R. (2020). Secure ear biometrics using circular kernel principal component analysis, Chebyshev transform hashing and Bose-Chaudhuri-Hocquenghem error-correcting codes. *Signal Image Video Proces.* 14, 847–855. doi: 10.1007/s11760-019-01609-y
- Osaki, Y., Kunin, M., Cohen, B., and Raphan, T. (2007). Three-dimensional kinematics and dynamics of the foot during walking: a model of central control mechanisms. *Exp. Brain Res.* 176, 476–496. doi: 10.1007/s00221-006-0633-1
- Osaki, Y., Kunin, M., Cohen, B., and Raphan, T. (2008). Relative contribution of walking velocity and stepping frequency to the neural control of locomotion. *Exp. Brain Res.* 185, 121–135. doi: 10.1007/s00221-007-1139-1
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). *Scikit-Learn: Machine Learning in Python* (Brookline, MA: Microtome Publishing), 2825–2830.
- Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Int. J. Machine Learn. Technol.* 2, 37–63. doi: 10.48550/arXiv.2010.16061
- Raphan, T. (2020). Vestibular, locomotor, and vestibulo-autonomic research: 50 years of collaboration with Bernard Cohen. *J. Neurophysiol.* 123, 329–345. doi: 10.1152/jn.00485.2019
- Saevanee, H., and Bhatarakosol, P. (2008). "User authentication using combination of behavioral biometrics over the touchpad acting like touch screen of mobile device," in *ICCEE 2008. International Conference on Computer and Electrical Engineering* (Phuket: ICCEE), doi: 10.1109/ICCEE.2008.157
- Sanchez, J., Mahmoudi, B., Digiovanna, J., and Principe, J. (2009). Exploiting co-adaptation for the design of symbiotic neural prosthetic assistance. *Neural Netw.* 22, 305–315. doi: 10.1016/j.neunet.2009.03.015
- Sky-News (2017). *Flight Diverted After Woman Unlocks Husband's Phone and Discovers Affair*. Sky News. Available online at: <https://news.sky.com/story/flight-diverted-after-woman-unlocks-husbands-phone-and-discovers-affair-11117184> (accessed October 25, 2020).
- Spillane, R. (1975). Keyboard apparatus for personal identification. *IBM Techn. Discl. Bull.* 1975, 17.
- Teh, P. S., Zhang, N., Teoh, A. B. J., and Chen, K. (2016). A survey on touch dynamics authentication in mobile devices. *Comput. Security* 59, 210–235. doi: 10.1016/j.cose.2016.03.003
- Telegraph, T. (2016). *The First Electric Telegraph in 1837 Revolutionised Communications*. The Telegraph. Available online at: <https://www.telegraph.co.uk/technology/connecting-britain/first-electric-telegraph/> (accessed January 17, 2021).
- Trewin, S., Swart, C., Koved, L., Martino, J., Singh, K., and Ben-David, S. (2012). "Biometric authentication on a mobile device: a study of user effort, error and task disruption," in *Proceedings of the 28th Annual Computer Security Applications Conference*. Orlando, FL: Association for Computing Machinery. doi: 10.1145/2420950.2420976
- Wang, X., Yu, T., Mengshoel, O., and Tague, P. (2017). "Towards continuous and passive authentication across mobile devices: an empirical study," in *Proceedings of the 10th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. Boston, MA: Association for Computing Machinery. doi: 10.1145/3098243.3098244
- Yiu, T. (2019). *The Curse of Dimensionality*. Towards Data Science. Available online at: <https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e?gi=c269056e0e89> (accessed January 5, 2021).
- Ziemke, T. (2003). "What's that thing called embodiment?," *Proceedings of Annual Meeting of the Cognitive Science Society* (Boston, MA: Cognitive Science Society), 25.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in

this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Jairam, Halevi and Raphan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

*The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*



# Machine Learning Solutions Applied to Amyotrophic Lateral Sclerosis Prognosis: A Review

Fabiano Papaiz<sup>1,2,3\*</sup>, Mario Emílio Teixeira Dourado Jr.<sup>1,4</sup>,  
Ricardo Alexsandro de Medeiros Valentim<sup>1</sup>, Antonio Higor Freire de Moraes<sup>1,3</sup> and  
Joel Perdiz Arrais<sup>2</sup>

<sup>1</sup> Laboratory of Technological Innovation in Health (LAIS), Federal University of Rio Grande do Norte, Natal, Brazil, <sup>2</sup> Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal, <sup>3</sup> Advanced Nucleus of Technological Innovation, Federal Institute of Rio Grande do Norte, Natal, Brazil, <sup>4</sup> Department of Internal Medicine, Federal University of Rio Grande do Norte, Natal, Brazil

## OPEN ACCESS

### Edited by:

Peter Kokol,  
University of Maribor, Slovenia

### Reviewed by:

Riccardo Rosati,  
Marche Polytechnic University, Italy  
Qian Du,  
GNS Healthcare, United States

### \*Correspondence:

Fabiano Papaiz  
fabianopapaiz@gmail.com

### Specialty section:

This article was submitted to  
Digital Public Health,  
a section of the journal  
Frontiers in Computer Science

**Received:** 03 February 2022

**Accepted:** 24 March 2022

**Published:** 28 April 2022

### Citation:

Papaiz F, Dourado MET Jr, Valentim  
RAM, Moraes AHF and Arrais JP  
(2022) Machine Learning Solutions  
Applied to Amyotrophic Lateral  
Sclerosis Prognosis: A Review.  
Front. Comput. Sci. 4:869140.  
doi: 10.3389/fcomp.2022.869140

The prognosis of Amyotrophic Lateral Sclerosis (ALS), a complex and rare disease, represents a challenging and essential task to better comprehend its progression and improve patients' quality of life. The use of Machine Learning (ML) techniques in healthcare has produced valuable contributions to the prognosis field. This article presents a systematic and critical review of primary studies that used ML applied to the ALS prognosis, searching for databases, relevant predictor biomarkers, the ML algorithms and techniques, and their outcomes. We focused on studies that analyzed biomarkers commonly present in the ALS disease clinical practice, such as demographic, clinical, laboratory, and imaging data. Hence, we investigate studies to provide an overview of solutions that can be applied to develop decision support systems and be used by a higher number of ALS clinical settings. The studies were retrieved from PubMed, Science Direct, IEEEExplore, and Web of Science databases. After completing the searching and screening process, 10 articles were selected to be analyzed and summarized. The studies evaluated and used different ML algorithms, techniques, datasets, sample sizes, biomarkers, and performance metrics. Based on the results, three distinct types of prediction were identified: Disease Progression, Survival Time, and Need for Support. The biomarkers identified as relevant in more than one study were the ALSFRS/ALSFRS-R, disease duration, Forced Vital Capacity, Body Mass Index, age at onset, and Creatinine. In general, the studies presented promissory results that can be applied in developing decision support systems. Besides, we discussed the open challenges, the limitations identified, and future research opportunities.

**Keywords:** Amyotrophic Lateral Sclerosis, prognosis, Machine Learning, health informatics, literature review

## 1. INTRODUCTION

Amyotrophic Lateral Sclerosis (ALS) is a rare, incurable, and progressive disease that affects the neurons of the human motor system. The communication between the brain and muscles is gradually interrupted, leading patients to paralysis and death. Its causes are unknown, typically commits men and women between the ages of 40 and 70. The average life expectancy is 3–5 years after symptoms onset, and the worldwide incidence is about 1.9 cases per 100,000 individuals



per year. ALS is clinically heterogeneous, presenting different sites of disease onset, extra-motor involvements, progression rates, and survival times among their patients (Andersen et al., 2012; Chiò et al., 2014; Swinnen and Robberecht, 2014; Hardiman et al., 2017). Being ALS a complex disease, providing an accurate prognosis becomes a challenge to the physicians (e.g., survival time, disease progression, moment to introducing specific treatments). Thus, it is essential to identify relevant biological markers (biomarkers) and understand how they are related to ALS disease progression. Biomarkers are parameters collected from the patients that can be used to confirm a disease presence (diagnosis), follow up a disease progression (prognosis) or treatment response (monitoring), and calculate the probability of developing a disease (risk) (Group, 2001). They can comprise different data types, such as clinical, biometric, imaging, biofluid, and genetic. Previous studies identified helpful biomarkers that can assist in ALS prognosis, such as age at symptom onset, diagnosis delay, weight loss, bulbar site of onset, rate of functional and respiratory impairment over time, microRNAs, neurofilaments, and laboratory tests (ALS, 1996; Cedarbaum et al., 1999; Kollwe et al., 2008; Chiò et al., 2009; Varghese et al., 2013; Hardiman et al., 2017; Waller et al., 2017).

Researches using Artificial Intelligence techniques, like Machine Learning (ML) algorithms, have been successfully applied to improve the diagnosis and prognosis of diseases, such as the recent advances in the oncology field (Kourou et al., 2015; O'Shea et al., 2016). The ML field aims to develop computer programs capable of learning using previous experience (training data) without being explicitly programmed for this. ML algorithms could extract information from the training data, transform it into knowledge, and use it to solve different categories of problems (e.g., classification, regression, clustering, Samuel, 1988). In theory, the greater the amount of training data available, the greater the algorithm's learning and performance (Mitchell, 1997; Kubat, 2017). In this sense, having access to ALS patient data is crucial to perform relevant studies in the prognostic area and create ML solutions to help physicians in their daily work. The analysis of medical data usually involves dealing with high-dimensional data, covering a large number of biomarkers. Thus, some ML techniques (e.g., Feature Selection, Dimensionality Reduction) can be applied to transform a complex dataset into a simpler one by identifying the more relevant biomarkers, which improve the learning performance, data collecting efficiency, and algorithm understanding (Lee and Verleysen, 2007; Brank et al., 2011). ML algorithms can be used to develop Clinical Decision Support Systems (CDSS). The CDSS are computer programs designed to help physicians make more appropriate and timely decisions about their patients (Berner et al., 2007; Beeler et al., 2014; Gultepe et al., 2014; CDS, 2015; Rosati et al., 2020; Romeo and Frontoni, 2022). These systems usually provide prognostic predictions to improve the decision-making process and, thus, improve the patient's quality of life. Some benefits include improving patients' quality of care, treatment efficiency, resource planning, and reducing costs. CDSS also represents a valuable tool to promote knowledge dissemination among all interested health workers. ML-based CDSS can improve clinical decisions

**TABLE 1 |** Research questions.

RQ	Question
01	What are the ALS databases used in the study?
02	How many patients comprise the cohort of the study?
03	What are the types of prediction addressed by the study?
04	What are the ML algorithms and techniques used in the study?
05	What are the biomarkers evaluated and the most relevant identified by the study?
06	What are the performances of the used ML algorithms?

by helping physicians analyze and make inferences on a large amount of patient data. However, some ML approaches present results that can not be easily understood, decreasing their interpretability (e.g., Artificial Neural Networks or Support Vector Machines). Interpretability refers to how well a person can understand the decisions made by the ML algorithm (Miller, 2019). This issue can difficult the process of acceptance and integration of a CDSS in the clinical environment routine. Consequently, the development of a CDSS must have concerned about interpretability issues, being transparent enough so that health workers can understand how any support was offered.

Many countries present financial limitations on their health system. This fact makes it unfeasible to collect complex and costly biomarkers (e.g., genetic) in primary care. In this manner, it is essential to carry out studies considering these limitations to develop computational solutions (e.g., CDSS) that can assist a higher number of primary care units.

The main objective of this study is to investigate ML approaches on ALS prognosis that analyzed less complex biomarkers, which can be potentially applied to develop clinical decision support systems to assist physicians in the real-world ALS clinical setting. We focused on studies that analyzed biomarkers commonly present in the ALS disease clinical practice, such as demographic, clinical (including functional, respiratory, and nutritional), laboratory, and imaging data. Hence, we investigate studies using biomarkers obtained through a less complex process, aiming to provide an overview of solutions that can be applied to develop decision support systems and be used on a large scale in primary care, considering financial limitations. In this sense, we did not include studies using *omics* data (i.e., genomic, transcriptomic, proteomic, and metabolomic). We described the recent advances in this area, the currently available datasets, the biomarkers analyzed, the ML algorithms and techniques used, the most relevant biomarkers identified, and their outcomes. Besides, we discussed the open challenges, the limitations identified, and future research opportunities.

## 2. METHODS

This systematic review aims to investigate ML solutions applied to ALS prognosis. In this sense, we elaborated research questions (RQ) to guide the conduct of this article, which are presented

**TABLE 2 |** Inclusion criteria.

IC	Description
01	Articles published in Journals
02	Articles written in English
03	Articles published between January 2011 and April 2021
04	Articles in the Information Technology, Computer Engineer, or Computer Science related areas

**TABLE 3 |** Exclusion criteria.

EC	Description
01	Review articles
02	Duplicate articles
03	Articles not related to Machine Learning applied to ALS prognosis
04	Articles using <i>omics</i> data (i.e., genomic, transcriptomic, proteomic, and metabolomic)

in **Table 1**. Next, we performed the following stages: (i) search articles related to ALS prognosis using ML in scientific databases, (ii) apply the inclusion criteria, (iii) apply the exclusion criteria, and (iv) analyze and summarize the selected articles.

In the first stage, the relevant literature was obtained from the *PubMed*, *Science Direct*, *IEEEExplore*, and *Web of Science* databases. The search was performed in April 2021 using the following search query: (“artificial intelligence” OR “machine learning” OR “deep learning”) AND (“amyotrophic lateral sclerosis” OR “motor neurone disease”) AND (“predict” OR “prognosis” OR “progression”). We used the Rayyan Web Application (Ouzzani et al., 2016) to organize the resulting articles and also to perform the remaining stages.

In the second and third stages, we applied the Inclusion (IC) and Exclusion (EC) Criteria to filter the articles according to the scope of this article (see **Tables 2, 3**). We considered only articles published in Journals, written in English, and published between January 2011 and April 2021 (IC-01, IC-02, and IC-03). Articles that did not belong to the Information Technology, Computer Engineer, or Computer Science related areas were not included (IC-04). Next, we carried out the removal of the review articles (EC-01), the duplicate entries (EC-02), and articles not related to ML applied to ALS prognosis (EC-03). Then, the articles using *omics* data were removed (EC-04).

Finally, in the fourth stage, the select articles were thoroughly read, which allowed the final analysis and accomplishment of the objectives of this research.

### 3. RESULTS

**Figure 1** illustrates the search and screening process for this systematic review. The search query and all inclusion criteria were used to perform the database searches. A total of 52 articles were retrieved, where two review articles were immediately excluded. After the removal of 15 duplicates, 35 articles were chosen for abstract review. A total of 25 studies were excluded

due to the use of *omic* data ( $n = 6$ ) and not being related to ML applied to ALS prognosis ( $n = 19$ ). After completing the searching and screening process, 10 articles were selected to be analyzed and summarized. The following sections present the results that address the research questions defined in this study (**Table 1**).

#### 3.1. ALS Datasets and Sample Sizes

Different datasets were analyzed and their sample sizes ranged from 41 up to over 10,000 samples. **Table 4** describes all the datasets analyzed. Most of the studies (60%) analyzed data from the PRO-ACT (Atassi et al., 2014) dataset, probably because it was the only publicly available. The other datasets used were local or proprietary. The data formats analyzed included tabular (all studies) and image (van der Burgh et al., 2017). More detail about the sample size used by each study are described in **Tables 6–8**.

#### 3.2. Types of Prediction Addressed

Based on the included studies, three distinct types of prediction were identified: *Disease Progression*, *Survival Time*, and *Need for Support* (more detail in **Table 5**). Kueffner et al. (2019) addressed the *Disease Progression* and *Survival Time* types simultaneously.

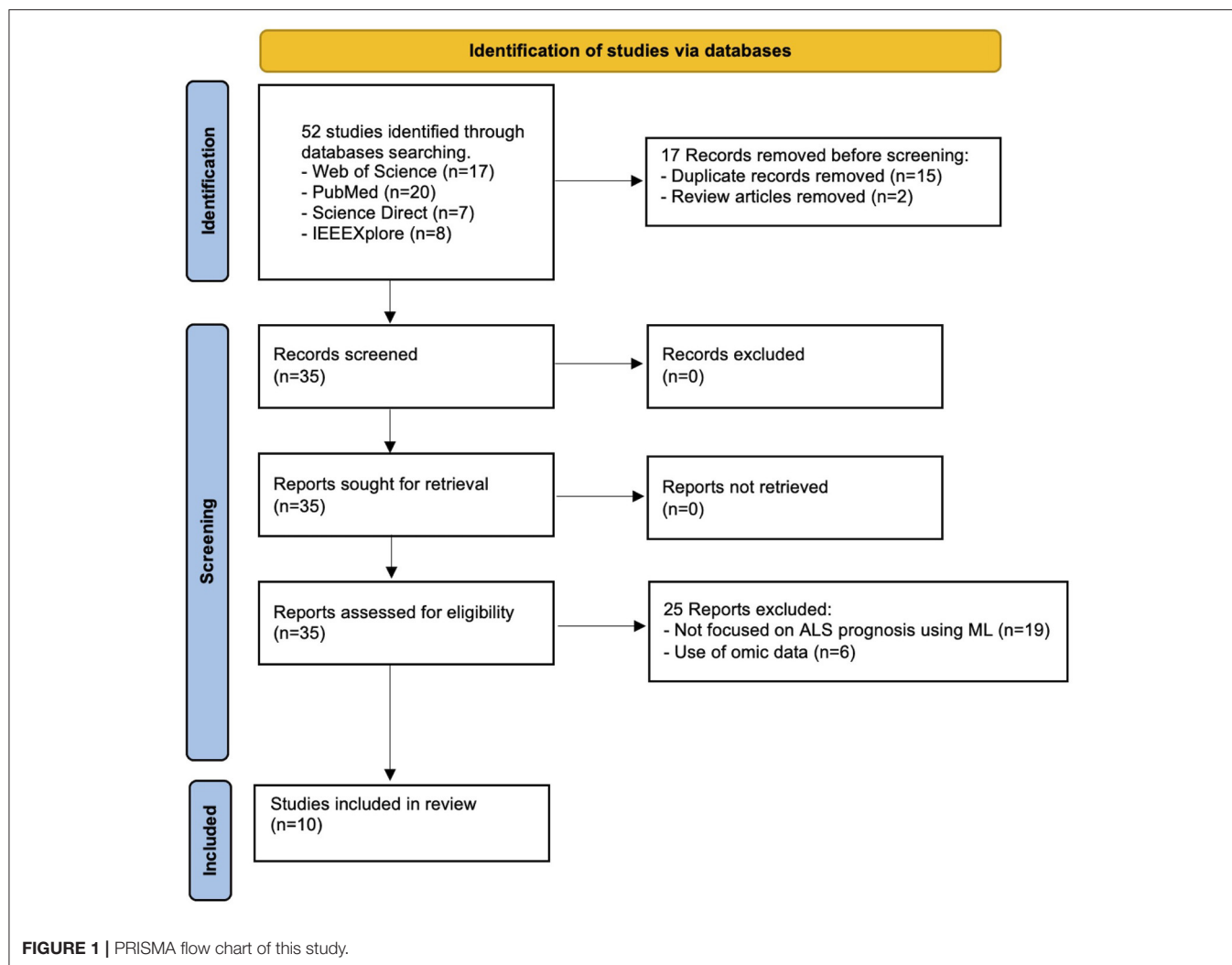
The *Disease Progression* prediction aimed to estimate the patient's state at a given moment in the future and was the type most addressed by the studies included (70%). The *Survival Time* prediction aimed to estimate the occurrence of death from a baseline date to a point-time in the future, such as the probability of death after 12 months from symptoms onset. The *Need for Support* prediction aimed to estimate the moment when patients will need more specialized support.

#### 3.3. Predictive Machine Learning Approaches

For *Disease Progression* prediction, most studies aimed to estimate changes in the ALS Functional Rating Scale (ALSFRS) or the Revised ALS Functional Rating Scale (ALSFRS-R) over time. Two other studies aimed to classify patients concerning their disease progression rates (Slow/Fast Kueffner et al., 2019, Low/High Greco et al., 2021). **Table 6** details the target predictions, best ML algorithm, performance, datasets, samples size, techniques, validation strategies, and biomarkers evaluated for each study.

The studies that addressed the *Survival Time* prediction aimed to classify the patients into survival groups and estimate the probability of death after a specific time interval. van der Burgh et al. (2017) aimed to classify patients into Short (<25 months), Medium (25–50 months), or Long (>50 months) survival groups. Kueffner et al. (2019) aimed to estimate the probability of survival after 12, 18, and 24 months. Grollemund et al. (2020) aimed to estimate the probability of patients being alive after 12 months. All three studies used the date of symptoms onset as the baseline date. The characteristics of each study are detailed in **Table 7**.

Pires et al. (2018) was the unique study that addressed the *Need for Support* prediction, aiming to estimate the need for Non-Invasive Ventilation (NIV) support after 3, 6, and 12 months. The characteristics of this study are detailed in **Table 8**.



### 3.4. Biomarkers Evaluated and the Most Relevant Identified

As previously mentioned, we focused on the biomarkers commonly present in the ALS disease clinical practice, being obtained in a less costly and complex way. The biomarkers evaluated comprise clinical, demographic, vital signs, respiratory, functional, laboratory, imaging, neurophysiological, and medication data. For more detail, please see column *Biomarkers Evaluated* in **Tables 6–8**. All the selected studies evaluated the ALS Functional Rating Scale (ALSFRS) or the Revised ALS Functional Rating Scale (ALSFRS-R) biomarkers. This fact highlights the importance of these biomarkers in monitoring ALS patients.

**Table 9** depicts the most relevant biomarkers identified in the studies, with the information about their associated types of prediction. They comprised clinical, imaging, functional, respiratory, and laboratory data. The biomarkers identified as relevant in more than one study were the ALSFRS/ALSFRS-R ( $n = 7$ ), disease duration ( $n = 5$ ), Forced vital capacity ( $n = 4$ ), Body mass index ( $n = 2$ ), age at onset ( $n = 2$ ), and Creatinine ( $n = 2$ ).

### 3.5. Description of the Studies

van der Burgh et al. (2017) demonstrated the positive impact of using Magnetic Resonance Images (MRI) along with clinical information to classify ALS patients into three survival groups: Short (<25 months), Medium (25–50 months), and Long (>50 months). The biomarkers evaluated were clinical information (e.g., site of onset, age at onset, ALSFRS slope, FVC) and MRI images (Structural Connectivity and Brain Morphology data) from 135 ALS patients. They developed Deep Neural Networks models and evaluate them in four scenarios using different biomarkers sets: (i) only Clinical Data, (ii) only Structural Connectivity MRI Data, (iii) only Brain Morphology MRI Data, and (iv) combining Clinical and MRI Data. The greater accuracy was obtained using the Clinical-MRI combined data (84%) compared to the other three strategies (Clinical: 69%; Structural Connectivity MRI: 63%; Brain Morphology MRI: 63%). They pointed out the power of Deep Neural Networks in making predictions using complex data. However, the relationships between input and output variables could not be easily

**TABLE 4 |** List of datasets used in the studies, inversely ordered by the sample size.

Dataset	Samples	References
PRO-ACT	+10,000	Gordon and Lerner, 2019; Halbersberg and Lerner, 2019; Kueffner et al., 2019; Tang et al., 2019; Grollemund et al., 2020; Hadad and Lerner, 2020
Ireland-Italia	1,479	Kueffner et al., 2019
Tel Aviv—Sourasky Medical Center	1,328	Hadad and Lerner, 2020
Lisbon—Saint Mary's Hospital	1,214	Pires et al., 2018; Leão et al., 2021
Paris—Tertiary referral Centre for ALS	646	Grollemund et al., 2020
Trophos Company	431	Grollemund et al., 2020
Exonhit Pharma	172	Grollemund et al., 2020
Utrecht—University Medical Center	135	van der Burgh et al., 2017
Italia	41	Greco et al., 2021

**TABLE 5 |** Types of prediction addressed by the studies.

Type	Number of studies	References
Disease progression	7	Gordon and Lerner, 2019; Halbersberg and Lerner, 2019; Kueffner et al., 2019; Tang et al., 2019; Hadad and Lerner, 2020; Greco et al., 2021; Leão et al., 2021
Survival time	3	van der Burgh et al., 2017; Kueffner et al., 2019; Grollemund et al., 2020
Need for support	1	Pires et al., 2018

recognized, needing more investigation to understand ALS progression better.

Pires et al. (2018) developed a model to predict when a patient will need NIV support according to a given time window (3, 6, and 12 months). They used the Portuguese ALS Dataset ( $n = 1,070$ ), combining the static and temporal data into a data structure called snapshot, which contains all information about a patient at a specific date. The patients were divided into three disease progression groups (Slow, Neutral, and Fast) and, for each group, their respective snapshots were used as learning instances to evaluate several ML models. A Feature Selection Ensemble approach was used to select the relevant biomarkers for each group. The Random Forest model obtained the best performance for 3, 6, and 12 months time window values. The relevant biomarkers present in all groups were BMI, FVC, and VC. Other relevant biomarkers (present in 75% of the time) were age at onset, disease duration, and ALSFRS score. The authors reported the advantage of using specialized ML models for different patient groups (e.g., disease progression groups) rather than create generalized models treating all the patients similarly.

Halbersberg and Lerner (2019) demonstrated the benefit of using temporal modeling, sequence clustering, and sequential pattern mining to predict the last patient state recorded

(ALSFRS score) based on his past information. To find relevant deterioration patterns in temporal patients data they developed a framework consisting of three stages: (i) group patients with similar progression using hierarchical clustering based on Dynamic Time Warping, (ii) perform pattern mining to find out common functional deterioration patterns among patients based on the SPADE sequence mining algorithm, and (iii) develop a Random Forest model to classify patients into their most similar cluster to predict their next disease state. The performance obtained by the proposed framework (Accuracy: 73, F1 score: 0.68, Mean Absolute Error: 0.3) was superior related to two other benchmark models (Random Forest and Long Short-Term Memory, both using no temporal modeling). They used static (e.g., age at onset, time from onset, gender) and longitudinal (ALSFRS scores and subscores) data of 2,590 subjects from the PRO-ACT dataset. The most important predictors reported were the previous ALSFRS score, the previous ALSFRS *Dressing* subscore, the previous *Climbing Stairs* subscore, the previous *Turning in Bed* subscore, the time from disease onset, and the deterioration pattern termed  $\langle E, G, I \rangle$  (i.e., a sequential declining in the *Writing*, *Dressing*, and *Walking* ALSFRS subscores).

Gordon and Lerner (2019) evaluated the capacity of ordinal classifiers to predict the functional decline of the patients. They used data about the first and last patient visits from the PRO-ACT dataset ( $n = 3,772$ ), analyzing the following biomarkers: clinical, demographic, ALSFRS, FVC, medication, vital signs, and laboratory tests. The target variables were all ten ALSFRS items (questions) separately. The patient states were mapped to the ALSFRS items, thus correlating patient state to disease progression for each point in time. Addressing the ordinal nature of the ALSFRS, they evaluated the following ordinal classifiers: Cumulative Link Models (CLM), Ordinal Decision Trees (ODT), and Cumulative Probability Tree (CPT). To evaluate their performances, they defined a penalizing system that accounts for various error severities differently. Thus, a classifier was less penalized when it predicted the value of 2 instead of 1 when the real value was 3. These three classifiers were compared with the Random Forest (RF), a non-ordinal classifier. The results showed that the CLM and ODT ordinal classifiers presented a similar performance and outperformed the RF classifier regarding the Mean Absolute Error measured in the best experiment scenario (CLM: 0.62–1.06; ODT: 0.63–1.01; RF: 1.01–1.61). For feature selection, the authors implemented an algorithm based on the  $J_3$  scattering matrix criterion for each ALSFRS item individually. The most relevant predictors were the FVC, the site of onset, the time from onset, and the laboratory tests Creatinine, CK, Chloride, Phosphorus, and Alkaline Phosphatase.

A crowdsourcing strategy was presented in Kueffner et al. (2019), where were selected 30 teams around the world to participate in an ALS stratification challenge. They asked the participants to create ML models to perform prediction tasks using the PRO-ACT and the Irish-Italian Registries datasets. The teams used patient data from the first three months and were limited to evaluate only six of all biomarkers available. The target predictions were the Disease Progression at 12 months (decline of the Functional Rate Scale) and the Probability of Survival at 12,



**TABLE 6 |** Overview of ML approaches on disease progression.

References	Target prediction	Best algorithm	Performance	Biomarkers evaluated	Dataset (Samples)	Techniques	Validation
Halbersberg and Lerner (2019)	Last patient state (ALSFRS score) recorded based on his past information	SPADE + DTW + Clustering Method	Accuracy: 73 F1 Score: 0.68 MAE: 0.30	<i>Tabular</i> : Clinical, demographic, laboratory, ALSFRS	PRO-ACT (2,590)		Hold-out
Gordon and Lerner (2019)	Last patient state (ALSFRS subscores) recorded based on his past information	CLM, and ODT	MAE (min-max): -CLM: 0.62–1.06 -ODT: 0.63–1.01	<i>Tabular</i> : Clinical, demographic, vital signs, laboratory, ALSFRS	PRO-ACT (3,772)	FS	10-Fold CV
Kueffner et al. (2019)	-ALSFRS score at 12 months, using data from the first 3 months and only 6 biomarkers. -Patients classification into slow/fast progression groups.	GBM, and RF	GBM (PRO-ACT): -Z-score: $\approx 12$ RF (Ireland-Italia) -Z-score: $\approx 6$	<i>Tabular</i> : Clinical, demographic, vital signs, laboratory, FVC, SVC, ALSFRS	PRO-ACT (10,723) Ireland-Italia (1,479)		Hold-out
Tang et al. (2019)	ALSFRS score and FVC at 12 months, using 1 <sup>st</sup> visit and 3-month data	BART (ALSFRS), and RF (FVC)	BART: - $R^2$ : 0.22 - RMSE: 0.55 - Corr: 0.47 RF: - $R^2$ : 0.68 - RMSE: 14.27 - Corr: 0.83	<i>Tabular</i> : Clinical, demographic, pulse, BMI, FVC, laboratory, Riluzole medication, ALSFRS	PRO-ACT (2,424)	FS MI	5-Fold CV
Hadad and Lerner (2020)	ALSFRS score at several time intervals, varying from 6 up to 24 months	XGBoost	RMSE: 2.65–5.57 MAE: 1.98–4.42	<i>Tabular</i> : Clinical, demographic, vital signs, FVC, laboratory, ALSFRS	PRO-ACT (3,171) Tel Aviv (1,328)	FIA	Hold-out
Greco et al. (2021)	Patients classification into low/high progression rates groups	SVM	Accuracy: 87.25	<i>Tabular</i> : Clinical, demographic, laboratory, ALSFRS-R	Italia (41)	FS	LOO CV
Leão et al. (2021)	Changes in the ALSFRS-R score and subscores (before and after NIV)	Extension of DBN	Accuracy: 74–88 Sensitivity: 57–95 AUC: 75–98	<i>Tabular</i> : Clinical, demographic, El Escorial, BMI, C9orf72, FVC, MIP, MEP, PNRA, ALSFRS, ALSFRS-R	Lisbon (1,214)	MI	5-Fold CV

ALSFRS, ALS functional rating scale; ALSFRS-R, revised ALS functional rating scale; NIV, non-invasive ventilation; BMI, body mass index; FVC, forced vital capacity; SVC, slow vital capacity; MIP, maximum inspiratory pressure; MEP, maximum expiratory pressure; PNRA, phrenic nerve response amplitude; SPADE, sequential pattern discovery using equivalence class; DTW, dynamic time warping; CLM, cumulative link models; ODT, ordinal decision trees; GBM, generalized boosting model; RF, random forest; BART, Bayesian additive regression tree; SVM, support vector machine; DBN, dynamic Bayesian network; AUC, area under the ROC curve; MAE, mean absolute error; MSPE, mean squared prediction error;  $R^2$ , coefficient of determination; RMSE, root mean square error; Corr, Pearson's correlation coefficient; FS, feature selection; FIA, feature importance analysis; MI, missing data imputation; CV, cross-validation; LOO, leave-one-out.

18, and 24 months. Regarding the survival prediction, one team outperformed the others significantly using a Gaussian Process Regression model, presenting a better approach in leading with the right-censored patient outcome (dead or trial dropout). The best models related to the disease progression prediction used the Generalized Boosting Model and the Random Forest algorithms. The more relevant biomarkers were disease duration, age at onset, site of onset, gender, weight, BMI, respiratory exams (FVC and SVC), laboratory tests (Creatinine and Segmented Neutrophils), and ALSFRS scores and subscores. Based on the relevant biomarkers chosen by the teams, the authors have identified four distinct patient groups: Slow Progressing, Fast Progressing, Early Stage, and Late Stage. The main biomarkers related to each group were also detailed in this study, where

the authors highlighted the importance of the ALSFRS Bulbar subscore (questions 1–3) in discriminating between groups.

Tang et al. (2019) addressed predictions in changing of the ALSFRS score and in the FVC percentage. They used static and longitudinal biomarkers from the PROC-ACT dataset ( $n = 2,424$ ), including only those patients with information about ALSFRS scores over time. The longitudinal data were transformed into signature vectors aggregating statistics values (minimum, median, maximum, and slope). Using data from the first visit and at the 3-month, the authors create models to predict the changes in the ALSFRS slope at 12-month. The evaluated models (Random Forest and Bayesian Additive Regression Tree) achieved modest results (Correlation: 0.47; RMSE: 0.55;  $R^2$ : 0.22), thus, indicating the difficulty in predicting

**TABLE 7 |** Overview of ML approaches on survival time.

References	Target prediction	Best algorithm	Performance	Biomarkers evaluated	Dataset (Samples)	Techniques	Validation
van der Burgh et al. (2017)	Patients classification into Short (<25 months), Medium (25–50), and Long (>50) survival groups	Deep neural networks	Accuracy: 84	<i>Tabular</i> : Clinical, demographic, C9orf72, FTD, El Escorial, ALSFRS. <i>Image</i> : MRI.	Utrecht (135)	FS MI	Hold-out
Kueffner et al. (2019)	Probability of death within 12, 18, and 24 months	Gaussian Regression	PRO-ACT: -Z-score: $\approx 14.5$ Ireland-Italia: -Z-score: $\approx 13$	<i>Tabular</i> : Clinical, demographic, vital signs, laboratory, FVC, SVC, ALSFRS	PRO-ACT (10,723) Ireland-Italia (1,479)		Hold-out
Grollemund et al. (2020)	1-year survival prediction, classifying patients into high, intermediate, and low survival rates groups	UMAP	BAcc: 91% F1 Score: 96%	<i>Tabular</i> : Clinical, demographic, ALSFRS	PRO-ACT (3971) Trophos (431) Exonhit (172) Paris (646)		Hold-out

ALSFRS, ALS functional rating scale; ALSFRS-R, revised ALS functional rating scale; MRI, magnetic resonance image; FTD, frontotemporal dementia; FVC, forced vital capacity; SVC, slow vital capacity; UMAP, uniform manifold approximation and projection; BAcc, balanced accuracy; FS, feature selection; MI, missing data imputation.

**TABLE 8 |** Overview of ML approach on need for support.

References	Target prediction	Best algorithm	Performance	Biomarkers evaluated	Dataset (Samples)	Techniques	Validation
Pires et al. (2018)	Patients need for NIV support at 3, 6, and 12 months for three progression groups (slow, neutral, and fast)	RF	Slow: (3/6/12 months) - AUC: 81/87/91 - Sens: 70/72/78 - Spec: 76/83/86 Neutral: (3/6/12 months) - AUC: 76/82/86 - Sens: 58/62/79 - Spec: 78/83/77 Fast: (3/6/12 months) - AUC: 72/81/79 - Sens: 51/71/74 - Spec: 77/76/71	<i>Tabular</i> : Clinical, demographic, El Escorial, BMI, C9orf72, VC, FVC, P0.1, SNIP, MIP, MEP, NIV, PNRA, PNRL, CE, CF, ALSFRS, ALSFRS-R	Lisbon (1070)	FS DB	10-Fold CV

ALSFRS, ALS functional rating scale; ALSFRS-R, revised ALS functional rating scale; NIV, non-invasive ventilation; BMI, body mass index; FVC, forced vital capacity; SVC, slow vital capacity; VC, vital capacity; P0.1, airway occlusion pressure; SNIP, sniff nasal inspiratory pressure; MIP, maximum inspiratory pressure; MEP, maximum expiratory pressure; PNRA, phrenic nerve response amplitude; PNRL, phrenic nerve response latency; CE, cervical extension; CF, cervical flexion; RF, random forest; AUC, area under the ROC curve; Sens, sensitivity; Spec, specificity; FS, feature selection; DB, data balancing; CV, cross-validation.

12-month ALSFRS slope using the only baseline and 3-months data. Feature Selection was performed using the Random Forest and the Knockoff Filter methods. After combining the top-ranked biomarkers returned by both methods, the best predictive biomarkers were the ALSFRS score, the disease duration, the FVC, and the Absolute Monocyte Count. To predict the FVC Percentage changes between 3 and 12 months, Random Forest models were tested in two scenarios (either including the baseline FVC or not). The best results were obtained using the FVC at baseline data, demonstrating the power of this biomarker, which increased the correlation from 0.67 to 0.83. The authors also applied unsupervised classification (K-Means) to find distinct phenotypes groups, founding four balanced clusters among the patients. However, it was considered impractical to clearly

understand how the groups differ due to the high number of biomarkers defined for each group during the clustering process.

Hadad and Lerner (2020) studied prediction of the ALSFRS score in several time intervals, varying from 6 to 24 months. Temporal (Long Short Term Memory—LSTM) and non-temporal (Random Forest, XGBoost, and Multilayer Perceptron) models were evaluated over the PRO-ACT dataset ( $n = 3,171$ ). To be used by the non-temporal models, the longitudinal data were transformed into vectors containing aggregated values (mean, standard deviation, slope, minimum, maximum). Each model was tested using 60 different randomly generated configurations, and their averaged performances were compared (Root Mean Square Error and Mean Absolute Error). The XGBoost model obtained superior performance for the most

**TABLE 9 |** Most relevant biomarkers identified, associated predictions, and references.

Type	Biomarker	Associated predictions/References		
		Disease progression	Survival time	Need for support
Clinical	Age at disease onset	Halbersberg and Lerner, 2019	Kueffner et al., 2019	–
	Body Mass Index (BMI)	Kueffner et al., 2019; Leão et al., 2021	Kueffner et al., 2019	–
	Disease duration	Gordon and Lerner, 2019; Halbersberg and Lerner, 2019; Kueffner et al., 2019; Tang et al., 2019; Leão et al., 2021	Kueffner et al., 2019	–
	Site of onset	Gordon and Lerner, 2019	–	–
Imaging	Magnetic Resonance Imaging	–	van der Burgh et al., 2017	–
Functional	ALSFRS	Halbersberg and Lerner, 2019; Kueffner et al., 2019; Tang et al., 2019; Hadad and Lerner, 2020	Kueffner et al., 2019; Grollemund et al., 2020	Pires et al., 2018
	ALSFRS-R	Leão et al., 2021	–	Pires et al., 2018
Respiratory	Forced Vital Capacity (FVC)	Gordon and Lerner, 2019; Kueffner et al., 2019; Tang et al., 2019	Kueffner et al., 2019	Pires et al., 2018
	Maximal expiratory pressure (MEP)	Leão et al., 2021	–	–
	Maximal inspiratory pressure (MIP)	Leão et al., 2021	–	–
	Slow vital capacity (SVC)	Kueffner et al., 2019	Kueffner et al., 2019	–
	Vital capacity (VC)	–	–	Pires et al., 2018
Laboratory	Absolute monocyte count	Tang et al., 2019	–	–
	Alanine transaminase (ALT)	Tang et al., 2019	–	–
	Alkaline phosphatase	Gordon and Lerner, 2019	–	–
	Calcium	Tang et al., 2019	–	–
	Chloride	Gordon and Lerner, 2019	–	–
	Cholesterol—Total	Greco et al., 2021	–	–
	Cholesterol—high-density (HDL)	Greco et al., 2021	–	–
	Creatine kinase (CK)	Gordon and Lerner, 2019	–	–
	Creatinine	Gordon and Lerner, 2019; Kueffner et al., 2019	–	–
	Hematocrit	Tang et al., 2019	–	–
	Phosphorus	Gordon and Lerner, 2019	–	–
	Potassium	Tang et al., 2019	–	–
	Segmented neutrophils	Kueffner et al., 2019	–	–
	Urine Ph	Kueffner et al., 2019	–	–
	Vitamin B12	Greco et al., 2021	–	–

time intervals evaluated (RMSE: 2.65–5.57, MAE: 1.98–4.42), being more precise for shorter than longer intervals. The relevant predictive biomarkers were the ALSFRS subscores. In another experiment, these models were evaluated in two scenarios: (i) trained with the PRO-ACT and tested with the TSMC dataset ( $n = 1,328$ ), and (ii) trained and tested using only the TSMC data. The short-term predictions (up to 6 months) were more precise using models trained with the PRO-ACT, and the XGBoost obtained the best results again. The authors highlighted that the PROC-ACT contains data from clinical trials that may not reflect the reality presented by the clinical environment patients due to the inclusion/exclusion criteria used. Thus, their patients tend to be younger and to have a slower disease progression, in addition to having more visits registered than the usual

clinical patients. To address this problem, they proposed a final experiment applying the Domain Adaptation approach to develop predictive models using the PRO-ACT data and improve their performances using patient clinical data. Firstly, LSTM and Multilayer Perceptron models were trained using only data from the PRO-ACT. Then, the training phase was complemented using the TSMC data to fine-tune the models to the clinical data. The results demonstrated that the use of domain adaptation improved the predictive performance for both models.

Grollemund et al. (2020) presented a dimensionality reduction model to predict 1-year survival rates. The biomarkers analyzed were gender, site onset, age, weight, disease duration, ALSFRS scores, ALSFRS slopes, and if died or not after one year. They

combined data from four datasets (PRO-ACT, Trophos, Exonhit, and Paris Tertiary Referral Center), totaling 5,220 samples. The obtained dataset was further divided into development and validation sets. After, the high-dimensional data from the development set were reduced and projected onto 2D space through the Uniform Manifold Approximation and Projection (UMAP) algorithm. Thus, the authors were able to project information about the patients into a 2D graph. The 2D data were divided into three 1-year survival probability zones: High (90%), Intermediate (80%), and Low (58%). Then, the validation set was used to evaluate the proposed model, and the results were compared with the Random Forest and the Logistic Regression models. The UMAP model obtained better classification results (F1 score: 96%, Balanced Accuracy: 91%) when compared to the average results of the other models (F1 score: 50%, Balanced Accuracy: 60%). The adopted approach also helped identify the biomarkers with higher or lower correlation with the survival prediction. For example, the age and ALSFRS score presented a high correlation, while the gender and weight showed a low correlation. However, the total comprehension of the relationship between input and output variables cannot be obtained because the adopted model is considered a black-box approach, which degrades its interpretability.

Despite Greco et al. (2021) aimed to find blood analytes to distinguish patients who have ALS from those with Lower Motor Neuron Disease (LMND), they also studied the classification of these patients with relation to their disease progression rates (High or Low). They analyzed clinic, demographic, and blood (108 analytes) data from 41 ALS patients. An SVM model was developed, and the Recursive-Feature-Elimination algorithm was used as a feature selection method. This model obtained an accuracy of 87.25% in classifying ALS patients into the High and Low groups using the first 16 ranked analytes, indicating the potential of using blood data as predictor biomarkers. Elevated levels of Vitamin-B12, Total Cholesterol, and HDL were related to a higher disease progression rate.

Leão et al. (2021) proposed a predictive model based on Dynamic Bayesian Networks (DBN), including both static and longitudinal data. They accessed data from the Portuguese ALS dataset ( $n = 1,214$ ), and the target prediction was the disease progression (ALSFRS score and subscores) related to the need for NIV support. To be processed by the DBN model, the longitudinal data were converted into time-series data and then divided into Before NIV and After NIV subsets. Thus, they were able to determine the most relevant biomarkers related to these two essential disease stages. The authors developed a predictive model, termed stdDBN framework, which uses stationary DBNs to predict disease progression and non-stationary DBNs to determine how the biomarkers analyzed change over time in each subset. The average results for predicting disease progression were above 80% for both subsets regarding the Accuracy, Sensitivity, and AUC metrics, demonstrating the potential of the proposed methodology. Graphs were generated to visualize how the biomarkers change over time, displaying their values in different time steps for each stage (before and after NIV). This approach allowed identifying some interesting relationships, as following mentioned. The Maximum Expiratory Pressure

(MEP) was considered the most important respiratory exam to predict the patient ventilatory decline before the need for NIV support. The ALSFRS Bulbar subscore had more influence on disease progression after NIV than before NIV. The BMI and Disease Duration had a stronger influence than the other static biomarkers for both subsets.

## 4. DISCUSSION

This study systematically reviewed the literature to identify relevant studies that used ML approaches to assist ALS disease prognosis. As explained before in Section 2, we focused on those studies comprising biomarkers commonly present in the daily ALS clinical practice. We identified 10 studies and detailed their target predictions, best ML algorithm, performance, datasets, samples size, techniques, validation strategies, biomarkers evaluated, and the most relevant biomarkers identified.

### 4.1. ALS Datasets and Data Preprocessing

Notably, the studies accessed datasets that concentrate ALS patients from Europe and the United States of America. Data from other regions were not analyzed (e.g., South America, Africa, or Asia). We consider this analysis essential to confirm (or not) if the predictive ML solutions can be broadly generalized and if different datasets can be combined to compose an even more relevant ALS dataset. Most of the studies (60%) analyzed data from the PRO-ACT dataset. PRO-ACT is the largest public ALS dataset available, containing over 10,000 samples, serving as a basis for several studies on ALS disease, and suitable for developing ML solutions. However, some studies included advised that the PRO-ACT has limitations that can increase the risk of creating biased models (Tang et al., 2019; Grollemund et al., 2020; Hadad and Lerner, 2020). Previous studies also reported these PRO-ACT limitations, and the risk of it does not represent the clinical patient population due to the inclusion and exclusion criteria used in the clinical trials (Chio et al., 2011; Atassi et al., 2014). For instance, their patients tend to be younger and present fewer functional impairments. In this sense, using a validation strategy that includes an external dataset represents an alternative to decrease bias risk and achieve a more reliable ML algorithm evaluation. This strategy was utilized by Hadad and Lerner (2020) and Grollemund et al. (2020). Hadad and Lerner (2020) created a training dataset combining samples from the PRO-ACT (100%) and Tel Aviv (90%) dataset. The samples remaining (10%) of the Tel Aviv dataset were used to test the model. Grollemund et al. (2020) performed the validation using the Paris dataset, which was not used in the training and testing stages. Preferably, the external dataset should contain data from the clinical patient population.

When designing ML solutions, we need to be aware of issues that can affect the performance and reliability of the model, such as missing values or data imbalance. The PRO-ACT dataset presented a considerable amount of missing values what caused that only 32% of its samples could be used in practice. Thus, it is valuable to evaluate how the missing data imputation methods can help to increase the sample size. van der Burgh et al. (2017) and Tang et al. (2019) used a more



straightforward imputation method, calculating the average for each feature and imputed it in the samples with missing values. Leão et al. (2021) combined the results of Last Observation Carried Forward and Linear Interpolation missing imputation methods, eliminating posteriorly the samples that still presented some missing values. However, the authors did not detail the sample sizes increase by using these strategies. The data imbalance problem occurs when the training data presents an unequal distribution between samples regarding some class of interest. Pires et al. (2018) combined Undersampling and Oversampling techniques to achieve a balance of 50% between the classes of interest. Grollemund et al. (2020) reported that the data imbalanced related to the target prediction (1-year survival probability) influenced the choice of adequate evaluation metrics due to 75% of the patients had survived for more than 1 year.

## 4.2. Predictive Biomarkers Analysis

Although some biomarkers evaluated are collected longitudinally (e.g., ALSFRS, respiratory, laboratory), most studies modeled these temporal data as non-temporal by summarizing longitudinal data into single values (e.g., slope, minimum, maximum, mean, standard deviation). This approach is termed Summary Measures and has some advantages such as being simple to comprehend, can be applied with unequal time intervals between measurements, and being considered statistically robust and valid (Matthews et al., 1990). It allowed that longitudinal information could be processed by non-temporal ML algorithms (e.g., Random Forest, XGBoost) to develop predictive solutions. However, this approach can hide some details about the biomarker changes over time because the aggregated value represents a linear variation over time. For example, an ALSFRS slope decline of 10 in 12 months can be seen as a decline of 0.84 per month (i.e., a linear decline), but the decline may have been accentuated only in the last three months. Future ALS prognosis studies can address this subject by comparing the results obtained using Summary Measures and longitudinal data, depicting the advantages and disadvantages of each approach. Approaches using temporal ML algorithms were presented by Halbersberg and Lerner (2019), Hadad and Lerner (2020), and Leão et al. (2021). Pires et al. (2018) used a strategy to create several snapshots representing the patient states over time by combining static and longitudinal data.

Regarding the ALSFRS/ALSFRS-R biomarker, we consider the approach of analyzing each subscore separately (e.g., swallowing, walking, writing, respiratory) should be preferred instead of analyzing the total score solely. A more precise analysis of the functional loss characteristics among patients can be performed. For example, two patients can have the same total score but with different values in their subscores, indicating a different disease progression for each patient. In the studies included, this approach helped to find distinct biomarkers associated with each subscore (Gordon and Lerner, 2019; Tang et al., 2019; Leão et al., 2021).

Different FS strategies were used by the studies included, which helped to find the more relevant biomarkers related to ALS disease (see **Table 9** for more detail). Some benefits reported were described hereafter. The FS strategy used by Greco et al. (2021) helped to select the 16 best predictors among 108 blood analytes (a reduction of 85%). Two laboratory tests (Chloride and Alkaline Phosphatase) were first associated with ALS progression due to the FS strategy used by Gordon and Lerner (2019).

## 4.3. Predictive Machine Learning Approaches

We identified three types of prediction addressed by the studies included (*Disease Progression, Survival Time, and Need for Support*). The studies evaluated and used different ML algorithms, techniques, datasets, sample sizes, biomarkers, and performance metrics. Consequently, a direct comparison of their performances is difficult, even within a specific type of prediction. In general, the results showed a considerable decrease in the predictive performance when using data from the first 3 months to predict long-term patient functional changes (e.g., at 12 or 24 months). Therefore, performing long-term predictions is still challenging due to ALS heterogeneity and complexity. The high accuracies reported by van der Burgh et al. (2017) (87.25%) and Greco et al. (2021) (84%) were overshadowed by the reduced number of samples analyzed (135 and 41, respectively), representing an elevated risk of model overfitting. Overfitting occurs when the algorithm presents good performance when using the training data but reduced performance when using the validation data, occurring a super adjust to the training data.

Both ML algorithms used by van der Burgh et al. (2017) (Deep Neural Networks) and Grollemund et al. (2020) (Dimensionality Reduction) presented interpretability issues by being considered black-box approaches. In these studies, the total comprehension of the relationship between input and output variables can not be easily explained. Physicians will desire to understand how the predictions were obtained to verify if they make sense and are trustworthy to be used for prognostication. The complexity of ALS disease makes a large number of biomarkers necessary to obtain good model performances. This fact also complicates the model interpretability when using black-box approaches. Thus, FS strategies can become an important allied to increase the model interpretability by reducing the number of biomarkers necessary. Some ML frameworks also can be explored to explain predictions obtained with black-box models, such as SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016). These frameworks are part of a recent research field termed Explainable Artificial Intelligence (XAI) (Adadi and Berrada, 2018).

Finally, the research efforts analyzed in this review, which used only biomarkers commonly present in the ALS clinical practice, demonstrated promissory results that can be applied in developing CDSS. Unexpectedly, only Gordon and Lerner (2019) reported the development of an information system based on their predictive approach and its deployment in an ALS clinical setting. This fact can indicate an absence of CDSS in the ALS prognostic area. Thus, the massive knowledge produced

is not used to build decision support systems effective to assist physicians in their daily work. It is an essential step to verify if the results obtained by the studies will be confirmed in a real-world clinical environment. As the results are confirmed, the CDSS will become more reliable to be used as a support tool by the physicians, even when black-box approaches have been utilized. From a practical point of view, a CDDS to assist the ALS prognosis could provide numerous valuable predictions. For example, based on the current patient disease progression rate, the system can inform how much a functional condition is estimated to decline in the following months (e.g., speech, respiratory, walking, swallowing). With this information, physicians could plan adequate treatment for the patient and determine if additional support will be needed (e.g., wheelchair, non-invasive ventilation, gastrostomy, cough assist machine). It could also be helpful to keep patients and families informed to better prepare themselves for the changes resulting from the worsening of the disease.

## 5. CONCLUSIONS

ALS is a devastating and incurable disease with no effective treatments, leading patients to death within 3–5 years from symptoms onset. Research efforts are essential to understand better the progression of this complex disease and improve patients' quality of life. This study reviewed relevant articles published between 2011 and 2021 that addressed the development of ML solutions to support the ALS prognosis.

The studies are promising, but some aspects need special attention. The datasets concentrated patients' data mainly from the USA and Europe. Thus, there is a need to collect and analyze data from other world regions to ensure that the ML solutions can be, in fact, generalized to all populations. When analyzing medical data, the Missing Values and Data Imbalance problems need to be addressed to avoid a negative impact on models' performance and reliability. The model interpretability issue is another important point to consider when using ML algorithms considered black-box, such as Neural Networks and Dimensionality Reduction. Despite the research advances, there is a probable lack of CDSS to assist the physicians in their daily work on ALS disease prognosis.

## REFERENCES

- Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052
- ALS (1996). The amyotrophic lateral sclerosis functional rating scale: assessment of activities of daily living in patients with amyotrophic lateral sclerosis. *Arch. Neurol.* 53, 141–147. doi: 10.1001/archneur.1996.00550020045014
- Andersen, P. M., Abrahams, S., Borasio, G. D., de Carvalho, M., Chio, A., Van Damme, P., et al. (2012). EFNS guidelines on the clinical management of amyotrophic lateral sclerosis (MALS)-revised report of an EFNS task force. *Eur. J. Neurol.* 19, 360–375. doi: 10.1111/j.1468-1331.2011.03501.x

## LIMITATIONS OF THIS STUDY

This research was limited in terms of scope as it did not cover studies that used more complex biomarkers, such as *omics* data (i.e., genomic, transcriptomic, proteomic, and metabolomic).

The reduced number of studies included ( $n = 10$ ) can increase the risk of bias. We used a simplistic search query based on keywords. Probably, the number of studies could be increased by using more advanced search options, such as MeSH tags or semantic search.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

FP, AM, and JA: conceptualization and methodology. FP, MD, RV, AM, and JA: validation and writing—review and editing. FP: investigation, data curation, and writing—original draft preparation. MD, RV, AM, and JA: supervision. RV: project administration and funding acquisition. All authors contributed to the article and approved the submitted version.

## FUNDING

The Brazilian Ministry of Health funded the present study through the Scientific and Technological Development Applied to ALS project, carried out by the Laboratory of Technological Innovation in Health (LAIS), of the Federal University of Rio Grande do Norte.

## ACKNOWLEDGMENTS

This research was carried out in cooperation agreement between Federal University of Rio Grande do Norte and University of Coimbra. We express our sincere gratitude to all leaders that was involved with this agreement.

- Atassi, N., Berry, J., Shui, A., Zach, N., Sherman, A., Sinani, E., et al. (2014). The PRO-ACT database: design, initial analyses, and predictive features. *Neurology* 83, 1719–1725. doi: 10.1212/WNL.0000000000000951
- Beeler, P. E., Bates, D. W., and Hug, B. L. (2014). *Clinical Decision Support Systems*. EMH Swiss Medical Publishers. doi: 10.4414/sm.w.2014.14073
- Berner, E. S., Hannah, K. J., and Ball, M. J. (Eds.). (2007). "Clinical decision support systems," in *Health Informatics* (New York, NY: Springer New York). doi: 10.1007/978-0-387-38319-4\_1
- Brank, J., Mladenović, D., Grobelnik, M., Liu, H., Mladenović, D., Flach, P. A., et al. (2011). "Feature selection," in *Encyclopedia of Machine Learning*, eds C. Sammut and G. I. Webb (Boston, MA: Springer US), 402–406. doi: 10.1007/978-0-387-30164-8\_306
- CDS (2015). *Practical Predictive Analytics and Decisioning Systems for Medicine*. Amsterdam: Elsevier.

- Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., et al. (1999). The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J. Neurol. Sci.* 169, 13–21. doi: 10.1016/S0022-510X(99)00210-5
- Chio, A., Canosa, A., Gallo, S., Cammarosano, S., Moglia, C., Fuda, G., et al. (2011). ALS clinical trials: do enrolled patients accurately represent the ALS population? *Neurology* 77, 1432–1437. doi: 10.1212/WNL.0b013e318232ab9b
- Chio, A., Logroscino, G., Hardiman, O., Swinger, R., Mitchell, D., Beghi, E., et al. (2009). Prognostic factors in ALS: a critical review. *Amyot. Lateral Scler.* 10, 310–323. doi: 10.3109/17482960802566824
- Chio, A., Pagani, M., Agosta, F., Calvo, A., Cistaro, A., and Filippi, M. (2014). Neuroimaging in amyotrophic lateral sclerosis: insights into structural and functional changes. *Lancet Neurol.* 13, 1228–1240. doi: 10.1016/S1474-4422(14)70167-X
- Gordon, J., and Lerner, B. (2019). Insights into amyotrophic lateral sclerosis from a machine learning perspective. *J. Clin. Med.* 8:1578. doi: 10.3390/jcm8101578
- Greco, A., Chiesa, M. R., Da Prato, I., Romanelli, A. M., Dolciotti, C., Cavallini, G., et al. (2021). Using blood data for the differential diagnosis and prognosis of motor neuron diseases: a new dataset for machine learning applications. *Sci. Rep.* 11:3371. doi: 10.1038/s41598-021-82940-8
- Grollemund, V., Chat, G. L., Secchi-Buhour, M.-S., Delbot, F., Pradat-Peyre, J.-F., Bede, P., et al. (2020). Development and validation of a 1-year survival prognosis estimation model for Amyotrophic Lateral Sclerosis using manifold learning algorithm UMAP. *Sci. Rep.* 10:13378. doi: 10.1038/s41598-020-70125-8
- Group, B. D. W. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Therap.* 69, 89–95. doi: 10.1067/mcp.2001.113989
- Gultepe, E., Green, J. P., Nguyen, H., Adams, J., Albertson, T., and Tagkopoulos, I. (2014). From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *J. Am. Med. Inform. Assoc.* 21, 315–325. doi: 10.1136/amiajnl-2013-001815
- Hadad, B., and Lerner, B. (2020). “Domain adaptation from clinical trials data to the tertiary care clinic—Application to ALS,” in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Miami), 539–544. doi: 10.1109/ICMLA51294.2020.00090
- Halbersberg, D., and Lerner, B. (2019). “Temporal modeling of deterioration patterns and clustering for disease prediction of ALS patients,” in *2019 18th IEEE International Conference on Machine Learning And Applications (ICMLA)* (Boca Raton, FL), 62–68. doi: 10.1109/ICMLA.2019.00019
- Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., et al. (2017). Amyotrophic lateral sclerosis. *Nat. Rev. Dis. Primers* 3:17071. doi: 10.1038/nrdp.2017.71
- Kollewe, K., Mauss, U., Krampfl, K., Petri, S., Dengler, R., and Mohammadi, B. (2008). ALSFRS-R score and its ratio: a useful predictor for ALS-progression. *J. Neurol. Sci.* 275, 69–73. doi: 10.1016/j.jns.2008.07.016
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005
- Kubat, M. (2017). *An Introduction to Machine Learning*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-63913-0
- Kueffner, R., Zach, N., Bronfeld, M., Norel, R., Atassi, N., Balagurusamy, V., et al. (2019). Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Sci. Rep.* 9:690. doi: 10.1038/s41598-018-36873-4
- Leão, T., Madeira, S. C., Gromicho, M., de Carvalho, M., and Carvalho, A. M. (2021). Learning dynamic Bayesian networks from time-dependent and time-independent data: unraveling disease progression in Amyotrophic Lateral Sclerosis. *J. Biomed. Informatics* 117:103730. doi: 10.1016/j.jbi.2021.103730
- Lee, J. A., and Verleysen, M. (2007). “High-dimensional data,” in *Nonlinear Dimensionality Reduction*, Series Title: Information Science and Statistics, eds J. A. Lee and M. Verleysen (New York, NY: Springer New York), 1–16. doi: 10.1007/978-0-387-39351-3\_1
- Lundberg, S. M., and Lee, S.-I. (2017). “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, Vol. 30, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc.).
- Matthews, J. N., Altman, D. G., Campbell, M. J., and Royston, P. (1990). Analysis of serial measurements in medical research. *BMJ* 300, 230–235. doi: 10.1136/bmj.300.6719.230
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007
- Mitchell, T. M. (1997). *Machine Learning*, 1st Edn. McGraw-Hill series in computer science (New York, NY: McGraw-Hill)
- O’Shea, K., Cameron, S. J., Lewis, K. E., Lu, C., and Mur, L. A. (2016). Metabolomic-based biomarker discovery for non-invasive lung cancer screening: a case study. *Biochim. Biophys. Acta* 1860(11 Pt B), 2682–2687. doi: 10.1016/j.bbagen.2016.07.007
- Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Syst. Rev.* 5:210. doi: 10.1186/s13643-016-0384-4
- Pires, S., Gromicho, M., Pinto, S., Carvalho, M., and Madeira, S. C. (2018). “Predicting non-invasive ventilation in ALS patients using stratified disease progression groups,” in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (Singapore), 748–757. doi: 10.1109/ICDMW.2018.00113
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why should i trust you?: explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16* (New York, NY: Association for Computing Machinery), 1135–1144. doi: 10.1145/2939672.2939778
- Romeo, L., and Frontoni, E. (2022). A Unified Hierarchical XGBoost model for classifying priorities for COVID-19 vaccination campaign. *Pattern Recogn.* 121:108197. doi: 10.1016/j.patcog.2021.108197
- Rosati, R., Romeo, L., Silvestri, S., Marcheggiani, F., Tiano, L., and Frontoni, E. (2020). Faster R-CNN approach for detection and quantification of DNA damage in comet assay images. *Comput. Biol. Med.* 123:103912. doi: 10.1016/j.combiomed.2020.103912
- Samuel, A. L. (1988). “Some studies in machine learning using the game of checkers,” in *Computer Games I*, ed D. N. L. Levy (New York, NY: Springer New York), 335–365. doi: 10.1007/978-1-4613-8716-9\_14
- Swinnen, B., and Robberecht, W. (2014). The phenotypic variability of amyotrophic lateral sclerosis. *Nat. Rev. Neurol.* 10, 661–670. doi: 10.1038/nrneuro.2014.184
- Tang, M., Gao, C., Goutman, S. A., Kalinin, A., Mukherjee, B., Guan, Y., et al. (2019). Model-based and model-free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering. *Neuroinformatics* 17, 407–421. doi: 10.1007/s12021-018-9406-9
- van der Burgh, H. K., Schmidt, R., Westeneng, H.-J., de Reus, M. A., van den Berg, L. H., and van den Heuvel, M. P. (2017). Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *Neuroimage Clin.* 13, 361–369. doi: 10.1016/j.nicl.2016.10.008
- Varghese, A. M., Sharma, A., Mishra, P., Vijayalakshmi, K., Harsha, H. C., Sathyaprabha, T. N., et al. (2013). Chitotriosidase - a putative biomarker for sporadic amyotrophic lateral sclerosis. *Clin. Proteom.* 10:19. doi: 10.1186/1559-0275-10-19
- Waller, R., Goodall, E. F., Milo, M., Cooper-Knock, J., Costa, M. D., Hobson, E., et al. (2017). Serum miRNAs MIR-206, 143-3p and 374b-5p as potential biomarkers for amyotrophic lateral sclerosis (ALS). *Neurobiol. Aging* 55, 123–131. doi: 10.1016/j.neurobiolaging.2017.03.027

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Papaiz, Dourado, Valentim, de Moraes and Arrais. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques

B. Shamreen Ahamed\*, Meenakshi Sumeet Arya and Auxilia Osvin Nancy V

Department of Computer Science and Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Chennai, India

## OPEN ACCESS

### Edited by:

Bai Xue,  
Institute of Software (CAS), China

### Reviewed by:

Bohua Zhan,  
Institute of Software (CAS), China  
Qiuye Wang,  
Institute of Software (CAS), China

### \*Correspondence:

B. Shamreen Ahamed  
shamu1502@gmail.com

### Specialty section:

This article was submitted to  
Theoretical Computer Science,  
a section of the journal  
Frontiers in Computer Science

**Received:** 14 December 2021

**Accepted:** 08 April 2022

**Published:** 10 May 2022

### Citation:

Ahamed BS, Arya MS and Nancy V AO (2022) Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques.  
Front. Comput. Sci. 4:835242.  
doi: 10.3389/fcomp.2022.835242

The technological advancements in today's healthcare sector have given rise to many innovations for disease prediction. Diabetes mellitus is one of the diseases that has been growing rapidly among people of different age groups; there are various reasons and causes involved. All these reasons are considered as different attributes for this study. To predict type-2 diabetes mellitus disease, various machine learning algorithms can be used. The objective of using the algorithm is to construct a predictive model to critically predict whether a person is affected by diabetes. The classifiers taken are logistic regression, XGBoost, gradient boosting, decision trees, ExtraTrees, random forest, and light gradient boosting machine (LGBM). The dataset used is PIMA Indian Dataset sourced from UC Irvine Repository. The performance of these algorithms is compared in reference to the accuracy obtained. The results obtained from these classifiers show that the LGBM classifier has the highest accuracy of 95.20% in comparison with the other algorithms.

**Keywords:** prediction, machine learning, classifiers, accuracy, comparison

## INTRODUCTION

Diabetes mellitus (DM) is considered as a chronic disease that has been affecting people of all age groups. The exact cause of the disease is still unknown. However, some of the factors or causes include age, family history, other relative diseases, pregnancy, fluctuating glucose levels, blood pressure, etc. (Dash et al., 2019). Diabetes is a disease that can be controlled under medication; however, a complete cure through medicines is not possible as of today. Diabetes can belong to one of the four broad categories, such as type-1, type-2, gestational diabetes, or prediabetes (Nibareke and Laassiri, 2020). There are some sub-types classified under these four categories as well. "Type-1 diabetes" is also known as "insulin-dependent diabetes," which occurs when the insulin release cell is damaged and unable to produce insulin (Martinsson et al., 2020). In "type-2" diabetes, adequate amount of insulin is not produced in the body (Wang et al., 2015). This commonly happens at an average above age of 40 years. The "gestational diabetes (GDM)" occurs mostly during pregnancy. The last one among the main four categories, "prediabetes," occurs when the blood sugar level is higher than normal but not as high as type-2 diabetes (Mujumdar and Vaidehi, 2019).

In the recent years, many researchers are using the concept of machine learning to predict the DM disease. Some of the commonly used algorithms include logistic regression (LR), XGBoost (XGB), gradient boosting (GB), decision trees (DTs), ExtraTrees, random forest (RF), and light gradient boosting machine (LGBM). Each classifier has its own advantages over the other classifiers (Prabha et al., 2021). However, the classifier that gives the highest accuracy is determined in implementation.



This study is divided into different sections as follows: Section Related Works represents the related works in DM. Section Theoretical Concepts of the Classifiers determines the theoretical concepts of the various algorithms used. Section Results and Discussion determines the architecture and implementation of the classifiers. Section Conclusion and Future Work explains the conclusions and future works of the study.

## RELATED WORKS

The following researchers have used the concept of machine learning for predicting DM disease.

Khaleel and Al-Bakry (2021) have created a model to detect whether a person is affected with DM disease. The concept of machine learning (ML) is used for the detection procedures. The PIMA dataset is used for the study. The algorithms used are LR, Naive bayes (NB), and K-nearest neighbour (KNN). The accuracy obtained are 94, 79, and 69% from these algorithms. The measures such as precision, recall, and F-measure are taken into consideration and LR is considered to produce the highest accuracy.

Ahmed et al. (2021) have used ML algorithms, namely, DT, KNN, NB, RF, GB, LR, and support vector machine (SVM) for predicting DM. Preprocessing techniques, such as label-encoding-normalization, are used to increase the accuracy. Two different datasets are used. One dataset provides the highest accuracy for SVM with 80.26% and for the second dataset, the highest accuracy is given by DT and RF with 96.81%.

Maniruzzaman et al. (2018) have used the ML technique based on risk-stratification is developed, optimized and evaluated. Features are optimized using six feature selection techniques. Then PIMA Indian diabetes dataset (PIDD) is used. The 10 different classifiers are used. Both RF selection and RF classification techniques yield an accuracy of 92.26%.

Kumari et al. (2021) have used two datasets including PIDD and breast cancer dataset, which were taken from the UC Irvine (UCI) Repository. Three ML classifiers are used for prediction. They are RF, LR, and Naive Bayes. The accuracy obtained is the highest for both datasets with a percentage of 79.08% for PIMA data and 97.27% for breast cancer data using soft voting classifier.

Tigga and Garg (2020) have developed a prediction model for DM disease. A dataset was collected for the study consisting of 952 instances and 18 attributes. The PIMA dataset was also used. The machine learning classifiers used are RF, LR, KNN, SVM, NB, and DT. The accuracy obtained was the highest for RF with a percentage of 94.10% for collected data and 75% for PIMA dataset.

Diwani and Sam (2014) have developed a prediction model using 10-fold-cross-validation on the training and testing data. The Waikato environment for knowledge analysis tool has been used along with Naive Bayes and DTs algorithm. The accuracy obtained is the highest for Naive Bayes with 76.30%.

Butt et al. (2021) have proposed a machine learning based approach for early-stage identification, classification, and prediction of diabetes disease. The PIMA Indian dataset has been used. The classifiers used are RF, multilayer perceptron (MLP) and LR. The accuracy obtained is highest for MLP with 87.26%.

## THEORETICAL CONCEPTS OF THE CLASSIFIERS

The various classifiers that are used is explained in the following sub-sections.

### Logistic Regression

It is a statistics-based model that uses logical function to develop a binary-dependent variable. The relationship between dependent and independent variables is estimated based on probabilities (Diwani and Sam, 2014). The dependent variable is categorical in this method. Mathematically it is expressed as follows (Kaur and Chhabra, 2014):

$$h_{\theta}(x) = P(Y = 1|X; \theta)$$

The probability that  $Y = 1$  given  $X$  which is given as “ $\theta$ ”

$$P(Y = 1 | X; \theta) + P(Y = 0 | X; \theta) = 1$$

### The XGBoost

It is the implementation of gradient boosted DTs that are created sequentially. An important feature is its weights. Each individual variable is assigned a particular weight that are given to the DTs to obtain the results (Butt et al., 2021). The prediction scores of each individual DT is given by

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i)$$

where the number of trees is denoted by  $k$ , the functional space is given as  $f$ , and the possible set available is given as  $F$  (Patil et al., 2019).

### Gradient Boosting

Many weak learners are combined into a predictive model typically in the form of DTs (Sehly and Mezher, 2020). It is mainly used when we want to decrease the bias error. A gradient-descent technique is chosen to obtain values of the coefficients (Posonia et al., 2020).

The loss function used is  $(y_1 - y_1')^2$ .  $y_1$  is the actual value and  $y_1'$  is the final predicted value by this model. So  $y_1'$  is replaced with  $G_n(X)$ , which represents the actual target (Ke et al., 2017). It is mathematically expressed as follows:

$$G_{n+1}(X) = G_n(X) + \gamma_n H_1(x, e_n)$$

$$L_1 = (y_1 - y_1')^2$$

$$L_1 = (Y - G_n(x))^2$$

### Decision Trees

It is a supervised-learning algorithm (Islam et al., 2020). It works with categorical and continuous input and output variables. It is used to represent whether it belongs to classification or regression procedures (Chen and Guestrin, 2016). The types of DTs are as follows: ID3, ID 4.5, CART, and CHAID. The measures used on DT are as follows: Entropy, Gini index, and

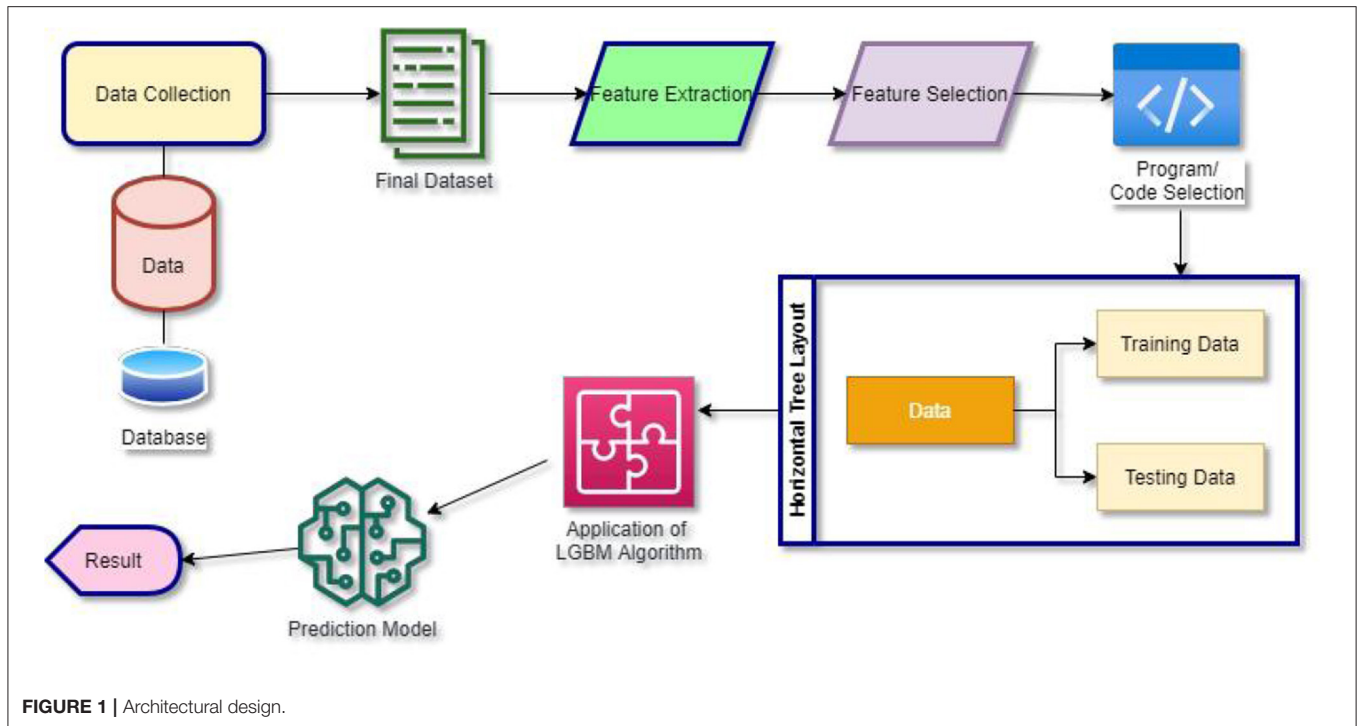


FIGURE 1 | Architectural design.

standard deviation (Khanam and Foo, 2021). It is mathematically calculated as follows (Ambigavathi and Sridharan, 2018):

$$\text{Entropy} = - \sum_{i=1}^{n1} p_{i1}^* \log(p_{i1})$$

$$\text{Gini Index} = 1 - \sum_{i=1}^{n1} p_{i1}^2$$

## Extra Trees

Extra trees (ETs) are also called as “extremely randomized trees classifier.” It is a type of “ensemble learning technique” which combines many decorrelated DTs to result as a single tree classification (Chen et al., 2017). It differs from RF in a way in which DTs are built. The entropy is calculated as follows:

$$\text{Entropy}(S1) = \sum_{i=1}^{c1} -p_{i1} \log_2 p_{i1}$$

where the number of unique class labels is given as  $c1$ , the proportion of rows with output label is given as  $p_{i1}$  (Sisodia and Sisodia, 2018).

Then the “information gain” is calculated using the following formula (Ke et al., 2017):

$$\text{Gain}(S1, A) = \text{Entropy}(S1) - \sum_{v \in \text{Values}(A)} \frac{|S1_v|}{|S1|} \text{Entropy}(S1_v)$$

## Random Forest

The RF combines the output of multiple DT to reach a single result. The DT is taken as a base and row sampling as well as

column sampling. The number of base learners is increased and the variance is decreased or *vice versa*. For cross-validation, K can be used. It is considered as an important bagging method (Mamuda and Sathasivam, 2017).

Random Forest = DT (base learner) + bagging (Row sampling with replacement) + feature bagging (column sampling) + aggregation (mean/median, majority vote)

## Light Gradient Boosting Machine

The performance of LGBM is considered to be high-performance and is represented as “GB framework” based on DT algorithm (Ahamed and Arya, 2021). It is majorly used for classifying and ranking. It splits the tree leaf-wise with best-fit. It can be measured using the data improvement technique and can be given by calculating the variance after segregating (Zhu et al., 2020). It can be represented as follows:

$$Y1 = \text{Base\_Tree}(X1) - lr1 * \text{Tree1}(X1) - lr1 * \text{Tree2}(X1) \dots$$

## System Architecture

The data needed for the study are initially collected and stored in the database. The dataset PIMA is taken from UCI Repository for execution. The dataset is then pre-processed using different exploratory data analysis techniques. The dataset is divided into “training data” and “testing data.” The various algorithms mentioned are then compared and the best working algorithm producing the highest accuracy is taken as the best predictive model for predicting DM disease. The architectural structure depicted in Figure 1.

**TABLE 1** | Accuracy percentage.

Dataset	Logistic regression	XGB classifier	Gradient boosting classifier	Decision tree	Extra trees classifier	Random forest	LGBM
PIMA Indian dataset	75.20%	83.30%	94.10%	94.40%	94.60%	94.80%	95.20%

## RESULTS AND DISCUSSION

The results and accuracy percentage calculated are given in the form of a table (Table 1).

The algorithms considered are LR, XGB, GB, DT, ET, RF, and LGBM. The accuracy obtained is the highest for LGBM with 95.2%.

## CONCLUSION AND FUTURE WORK

These discussions here were considered and we identified that “LGBM algorithm” worked best for the dataset taken by producing an accuracy that was higher in comparisons with the other algorithms. However, in future, different dataset can be taken and compared with the different classifiers to classify

which algorithm can produce the best result. Also, the parameters using in LGBM can be further finetuned and an advanced LGBM algorithm can be used and the prediction accuracy percentage can be increased.

## AUTHOR CONTRIBUTIONS

BA and MA: material preparation, data collection, analysis, resources, and writing—review and editing. BA: first draft of the manuscript and investigation. MA: conceptualization, supervision, and visualization. AN: coding and idea of research. All authors contributed to the study conception and design, involved in the idea for the article, performed the literature search, data analysis, drafted, and critically revised the work.

## REFERENCES

- Ahamed, B. S., and Arya, M. S. (2021). Prediction of Type-2 diabetes using the LGBM classifier methods and techniques. *Turk. J. Comput. Math. Educ.* 12, 223–231. Available online at: <https://www.proquest.com/docview/2622815314>
- Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., et al. (2021). Machine learning based diabetes prediction and development of smart web application. *Int. J. Cogn. Comp. Eng.* 2, 229–241. doi: 10.1016/j.ijcce.2021.12.001
- Ambigavathi, M., and Sridharan, D. (2018). “Big data analytics in healthcare,” in *IEEE Tenth International Conference on Advanced Computing (ICoAC)*, 269–276. doi: 10.1109/ICoAC44903.2018.8939061
- Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., and Sherazi, H. H. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *J. Healthc. Eng.* 2021, 9930985. doi: 10.1155/2021/9930985
- Chen, T., and Guestrin, C. (2016). “XGBoost: a scalable tree boosting system,” in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. doi: 10.1145/2939672.2939785
- Chen, W., Chen, S., Zhang, J. H., and Wu, T. (2017). “A hybrid prediction model for type 2 diabetes using K-means and decision tree,” in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 386–390. doi: 10.1109/ICSESS.2017.8342938
- Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *J. Big Data* 6, 54. doi: 10.1186/s40537-019-0217-0
- Diwani, S. A., and Sam, A. (2014). Diabetes forecasting using supervised learning techniques. *Adv. Comp. Sci. Int. J.* 3, 10–18. Available online at: <http://www.acsij.org/acsij/article/view/156>
- Islam, M. S., Qaraqe, M. K., Abbas, H. T., Erraguntla, M., and Abdul-Ghani, M. (2020). “The prediction of diabetes development: a machine learning framework,” in *2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering, MECBME 2020* (IEEE Computer Society). doi: 10.1109/MECBME47393.2020.9292043
- Kaur, G., and Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. *Int. J. Comp. Appl.* 98, 13–17. doi: 10.5120/17314-7433
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). “LightGBM: a highly efficient gradient boosting decision tree,” in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.
- Khaleel, F. A., and Al-Bakry, A. M. (2021). Diagnosis of diabetes using machine learning algorithms. *Mater. Today Proc.* doi: 10.1016/j.matpr.2021.07.196
- Khanam, J. J., and Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Exp.* 7, 432–439. doi: 10.1016/j.ict.2021.02.004
- Kumari, S., Kumar, D., and Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int. J. Cogn. Comp. Eng.* 2, 40–46. doi: 10.1016/j.ijcce.2021.01.001
- Mamuda, M., and Sathasivam, S. (2017). “Predicting the survival of diabetes using neural network,” in *Proceedings of the AIP Conference Proceedings* (Bydgoszcz), 40–46. doi: 10.1063/1.4995878
- Maniruzzaman, M., Rahman, M., Al-MehediHasan, M., Suri, H. S., Abedin, M., El-Baz, A., et al. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *J. Med. Syst.* 42, 92. doi: 10.1007/s10916-018-0940-7
- Martinsson, J., Schliep, A., Eliasson, B., and Mogren, O. (2020). Blood glucose prediction with variance estimation using recurrent neural networks. *J. Healthc. Inform. Res.* 4, 1–18. doi: 10.1007/s41666-019-00059-y
- Mujumdar, A., and Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Proc. Comp. Sci.* 165, 292–299. doi: 10.1016/j.procs.2020.01.047
- Nibareke, T., and Laassiri, J. (2020). Using big data-machine learning models for diabetes prediction and flight delays analytics. *J. Big Data* 7, 78. doi: 10.1186/s40537-020-00355-0
- Patil, M. K., Sawarkar, S. D., and Narwane, M. S. (2019). Designing a model to detect diabetes using machine learning. *Int. J. Eng. Res. Technol.* 8, 333–340. Available online at: <https://www.ijert.org/designing-a-model-to-detect-diabetes-using-machine-learning>
- Posonia, A. M., Vigneshwari, S., and Rani, D. J. (2020). “Machine learning based diabetes prediction using decision tree J48,” in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 498–502. doi: 10.1109/ICISS49785.2020.9316001

- Prabha, A., Yadav, J., Rani, A., and Singh, V. (2021). Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier. *Comp. Biol. Med.* 136, 104664. doi: 10.1016/j.combiomed.2021.104664
- Sehly, R., and Mezher, M. (2020). "Comparative analysis of classification models for pima dataset," in *International Conference on Computing and Information Technology (ICCIT-1441)*, 1–5. doi: 10.1109/ICCIT-144147971.2020.9213821
- Sisodia, D., and Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Proc. Comp. Sci.* 132, 1578–1585. doi: 10.1016/j.procs.2018.05.122
- Tigga, N. P., and Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Proc. Comp. Sci.* 167, 706–716. doi: 10.1016/j.procs.2020.03.336
- Wang, F., Stiglic, G., Obradovic, Z., and Davidson, I. (2015). Guest editorial: special issue on data mining for medicine and healthcare. *Data Min. Knowl. Disc.* 29, 867–870. doi: 10.1007/s10618-015-0414-1
- Zhu, T., Li, K., Chen, J., Herrero, P., and Georgiou, P. (2020). Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *J. Healthc. Inform. Res.* 4, 308–324. doi: 10.1007/s41666-020-00068-2

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ahamed, Arya and Nancy V. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Micro-HBI: Human-Biology Interaction With Living Cells, Viruses, and Molecules

Seung Ah Lee<sup>1†</sup> and Ingmar H. Riedel-Kruse<sup>2†</sup>

<sup>1</sup> Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, <sup>2</sup> Departments of Molecular and Cellular Biology and Applied Mathematics and Biomedical Engineering, University of Arizona, Tucson, AZ, United States

## OPEN ACCESS

### Edited by:

Stacey Kuznetsov,  
Arizona State University, United States

### Reviewed by:

Raphael Kim,  
Delft University of Technology,  
Netherlands

### \*Correspondence:

Seung Ah Lee  
seungahlee@yonsei.ac.kr  
Ingmar H. Riedel-Kruse  
ingmar@arizona.edu

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Computer Science,  
a section of the journal  
Frontiers in Computer Science

**Received:** 06 January 2022

**Accepted:** 06 April 2022

**Published:** 16 May 2022

### Citation:

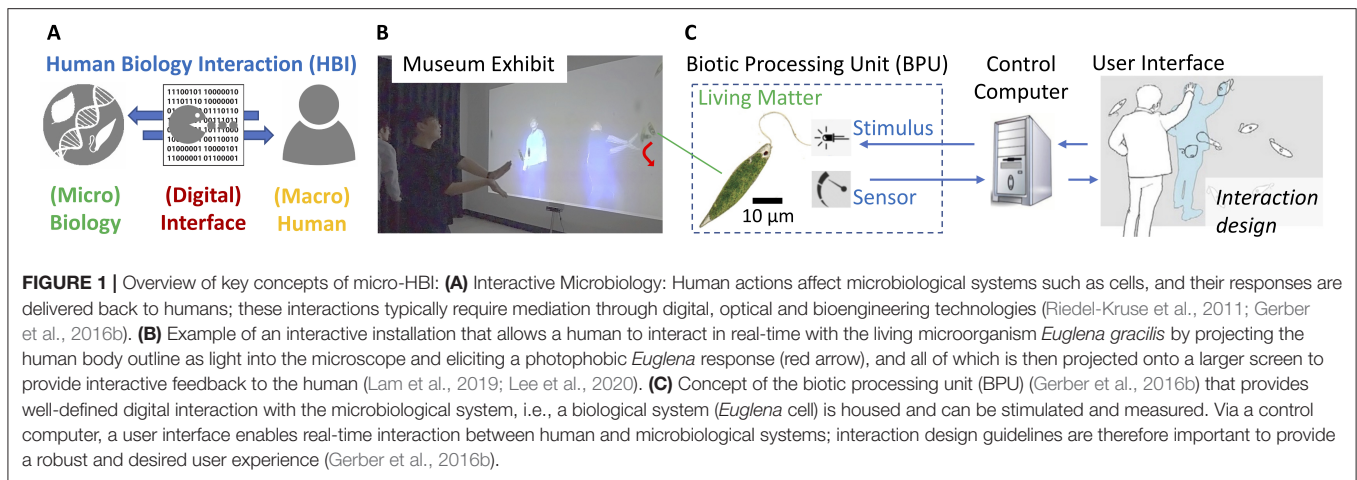
Lee SA and Riedel-Kruse IH (2022)  
Micro-HBI: Human-Biology Interaction  
With Living Cells, Viruses, and  
Molecules.  
Front. Comput. Sci. 4:849887.  
doi: 10.3389/fcomp.2022.849887

Human-Biology Interaction (HBI) is a field that aims to provide first-hand experience with living matter and the modern life-sciences to the lay public. Advances in optical, bioengineering, and digital technologies as well as interaction design now also enable real and direct experiences at the microscale, such as with living cells and molecules, motivating the sub-field of “micro-HBI.” This is distinct from simulating any biological processes. There is a significant need for HBI as new educational modalities are required to enable all strata of society to become informed about new technologies and biology in general, as we face challenges like global pandemics, environmental loss, and species extinctions. Here we review this field in order to provide a jump-off point for future work and to bring stakeholder from different disciplines together. By now, the field has explored and demonstrated many such interactive systems, the use of different microorganisms, new interaction design principles, and versatile applications, such as museum exhibits, biotic games, educational cloud labs, citizen science platforms, and hands-on do-it-yourself (DIY) Bio maker activities. We close with key open questions for the field to move forward.

**Keywords:** Human-Biology Interaction (HBI), microbiology, cloud labs, biotic games, human-computer interaction design, bioengineering, cells

## INTRODUCTION AND CONCEPTS

Humans exist in a close relationship with nature, and various inter-species interactions with macroscopic animals and plants have long enriched our lives. Meanwhile, modern life sciences and biotechnology are dramatically transforming our society—similar to the impact of information technology over the past few decades (Riedel-Kruse et al., 2011). These advancements provide new opportunities for humans to access living systems such as cells at the microscopic scale. Moreover, such first-hand experiences are also needed to inform the general public about biotechnology, life-sciences, medicine and ecology, and the current COVID-19 pandemic or loss of terrestrial biodiversity (Riedel-Kruse et al., 2011; Lockee, 2021). Hence **Human-Biology Interaction (HBI)** (**Figure 1A**) has emerged as a new interdisciplinary field (Lee et al., 2015a), conceptually related to Human-Computer Interaction (HCI), which aims to provide the general public with first-hand experiences of living matter (**Figure 1B**). While interactions with macroscopic living organisms (e.g., plants and animals) is often straightforward and requires minimal technology, interactions between humans and microscopic living matter such as protists, bacteria, viruses and DNA is accomplished with the help of modern optical, biological,



and digital technologies that facilitate inter-scale and cross-modality interactions. These conceptual and technical aspects merit particular attention and have therefore motivated the sub-field of “**micro-HBI**.” Note that these works are distinctly different from simulations and virtual realities (Loparev et al., 2017), as the presence and interaction of truly living biological material is at its core (Riedel-Kruse et al., 2011), while concepts of augmented and mixed reality are incorporated (Lam et al., 2020).

The early works in the domain stem from a longstanding **BioArt** tradition, where many artworks were created from the artist’s interaction with microscopic biological matter like bacteria (Osthoff, 2001) or cultured cells (Bakkum et al., 2007). In addition, in the HCI field, the concept of living media (Cheok et al., 2008) was introduced to augment digital systems by using biological matter as a user-interface modality. Then, direct and real-time interactions between human users and microorganisms and/or biochemical processes were proposed and also demonstrated in the form of **biotic games** for entertainment and education (Riedel-Kruse et al., 2011). In the meantime, citizen science platforms such as EteRNA enabled playful interactions with RNA folding questions while also providing experimental feedback within a few days (Lee et al., 2014; Koepnick et al., 2019). During the development of micro-HBI, the framework of **biotic processing units (BPU)** (Figure 1C) enabled more flexible and robust engineering and design for exploring various applications such as museum exhibits (Lee et al., 2015a) and cloud labs (Hossain et al., 2015), eventually leading to a first set of best practices regarding HBI design, technology and ethics (Harvey et al., 2014; Gerber et al., 2016b).

Since then, the field has been growing rapidly with an increasing number of stakeholders from various disciplines. HBI systems for various purposes have been pursued including, but not limited to, education (Hossain et al., 2016; Washington et al., 2019), entertainment (Kim et al., 2018a; van Eck and Lamers, 2018), art and installations (Kuznetsov et al., 2018; Lam et al., 2019; Lee et al., 2020) and as a new modality for interaction (Alistar and Pevero, 2020; Merritt et al., 2020; Pataranutaporn et al., 2020; Ofer et al., 2021). In addition, dedicated user studies

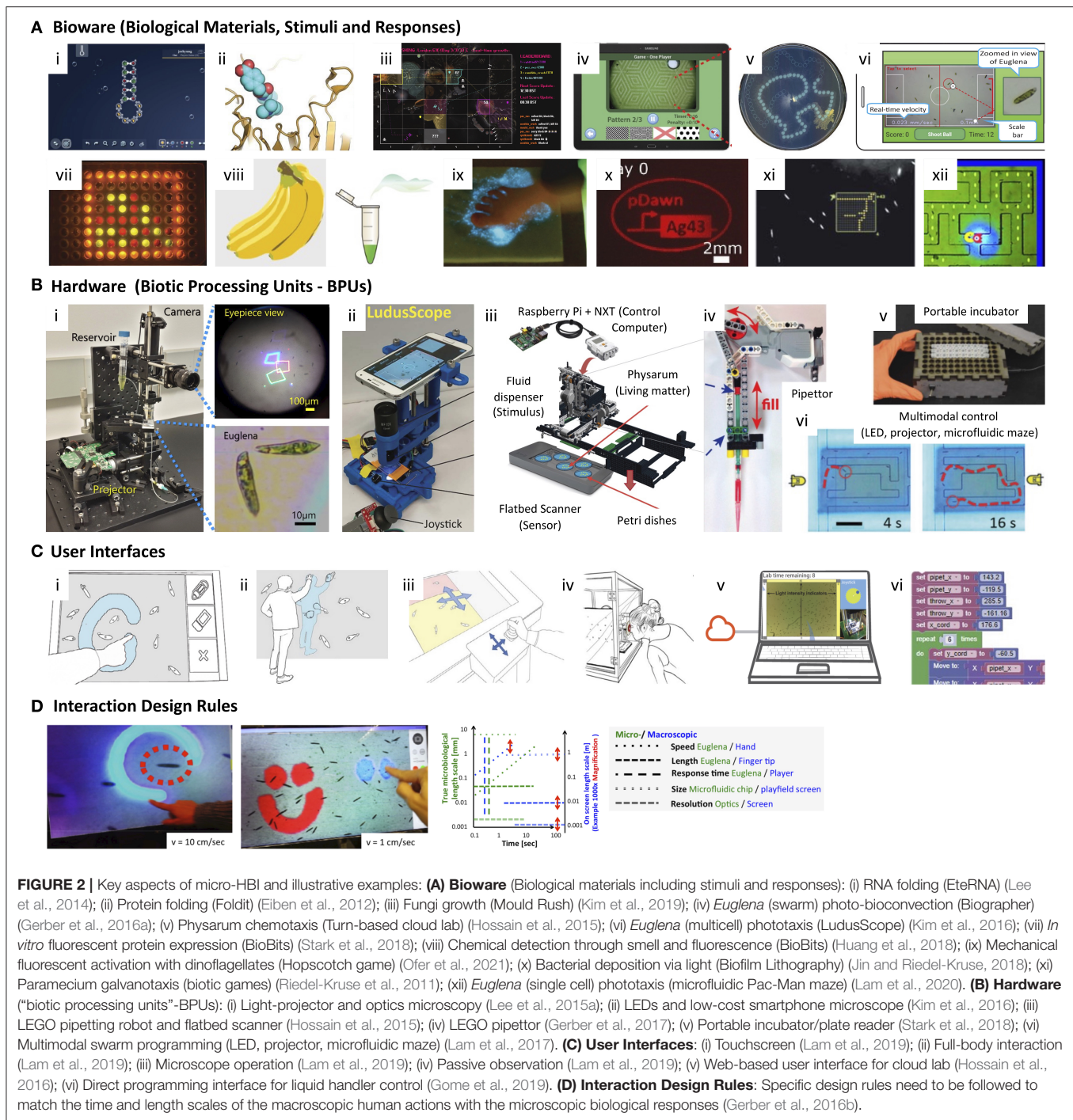
for HBI have been undertaken (Hossain et al., 2017b; Lam et al., 2019), providing more insight into the design of micro-HBI systems. We also witnessed a number of papers proposing related and overlapping concepts and terminology (some supported by practical work, while others were rather conceptual), for example, “Internet of Biotic Things (IoBT)” (Kim et al., 2018b), “Living Bits” (Pataranutaporn et al., 2020), “BioBits” (Huang et al., 2018; Stark et al., 2018), interactive biodesign (Gough et al., 2021), empathetic living media (Cheok et al., 2008) and living media interfaces (Merritt et al., 2020).

Due to the interdisciplinary nature of the field, the growth has often been horizontal rather than building upon each other. The field is becoming rich in concepts and speculative designs, making it difficult to obtain a comprehensive view of what has already been previously accomplished. In addition, advances in enabling technologies, design methodologies based on thorough user studies, and practical applications and use-cases are comparably overlooked. In this short review, we aim to outline the key concepts of micro-HBI, and we introduce and highlight some noteworthy work in the field as identified through an in-depth literature search. To move the field forward, we focus this review on the enabling technologies, design insights and ethical guidelines, ultimately suggesting future directions.

## NOTABLE PRIOR WORK

To list and introduce notable prior works in micro-HBI, we categorized them based on the application types and the end goals (Figure 2). We identified at least four major (partially overlapping) applications areas, i.e., education and scientific inquiry, art and play, human-computer interactions, and living and programmable materials.

The integration of real biological experiments and modeling for educational and scientific inquiry purposes was achieved with real-time interactive and turn-based **cloud labs** (Hossain et al., 2015, 2016). These systems were accessed in K-12 and university class-room settings as well as through massive open online courses (MOOCs) (Hossain et al., 2016, 2017a;



Hossain and Riedel-Kruse, 2018) for their **educational** outcomes. In hands-on settings, such activities can also be viewed as augmented, interactive microscopy (Kim et al., 2016), contrasting the traditional passive (including the more recent virtual) educational microscopy (Wilson et al., 2016). The citizen science games such as EteRNA and Foldit enabled folding predictions of RNA and protein molecules with real experimental feedback (Lee et al., 2014; Das et al., 2019). Do-it-yourself biology (DIY bio), synthetic biology such as "BioBits" and

low-cost equipment provide accessible platforms for education, engineering biological matter and artistic expression (Kuznetsov et al., 2012; Huang et al., 2018; Stark et al., 2018).

Various artistic and playful applications have been demonstrated and others have been proposed ("conceptual design"), many of which have also overlapped the same system or even within the same activity: **Open-ended play** and **rule-based play** (Riedel-Kruse et al., 2011; Lee et al., 2015a; Kim et al., 2016; Lam et al., 2020), **museum or public installations** (Lee et al.,



2015a, 2020; Lam et al., 2019), bioart (Kuznetsov et al., 2018), and sonification of microbiological behavior (Riedel-Kruse et al., 2011).

In the HCI field, many digital systems also continued to employ **microbes as a part of the interface media** (Alistar and Pevere, 2020; Merritt et al., 2020). In these examples, the major purpose of living media was to reinforce psychological factors and enrich the human user's experience. Cheok et al. (2008) showed that living matter can promote human empathy during engagement with digital systems, and Alistar and Pevere (2020) proposed using bacteria as tangible media to deliver information using human senses such as smell and touch. Beyond the interface media, efforts to develop digital systems have focused on the inter-species interaction between humans and microbes for ecological, educational and entertainment purposes (Chen et al., 2021; Ofer et al., 2021). These systems take the non-anthropocentric design approach that includes microbial life as an important stakeholder in the system rather than as a component of the system to be controlled and utilized (Lee et al., 2020).

Many other applications have been considered, such as interactive **biocomputation**, **swarm robotics** (Lam et al., 2017; Washington et al., 2019) and sensing, which further connect to the fields of hybrid living and **living smart materials** (Pataranutaporn and Lyle, 2018; Smith et al., 2020) and where self-assembly (Glass and Riedel-Kruse, 2018) or biofilm lithography (Jin and Riedel-Kruse, 2018) of adhesive cells may provide micro-LEGO bricks. Additionally, a wide range of more macroscopic applications exist such as **wearables and clothing** (Yao et al., 2015; Nguyen et al., 2021) as well as hybrid living materials (Smith et al., 2020). Other speculative future applications include **food and health applications** (Huang et al., 2018; Stark et al., 2018).

## MICROBIOLOGY, TECHNOLOGIES, AND INTERFACES

The design of interactive microbiology systems needs to focus on at least three major components, i.e., the biological subject matter ("bioware") (Figure 2A), the biotic processing units that both maintain and interface with the biological matters ("hardware") (Figure 2B), and the interaction modalities and interfaces for humans ("user interface") (Figure 2C).

Different types of microscopic materials and stimulus modalities ("bioware," Figure 2A) have been demonstrated (Gerber et al., 2016b). For example, single-celled **organisms** such as *Euglena* (Hossain et al., 2016), *paramecia* (Riedel-Kruse et al., 2011), *dinoflagellates* (Ofer et al., 2021) and bacteria (Alistar and Pevere, 2020; Chen et al., 2021), multicelled organisms such as the slimemold *physarum* (Hossain et al., 2015), molecules like RNA and DNA (Stojanovic and Stefanovic, 2003; Riedel-Kruse et al., 2011; Lee et al., 2014), cell collectives (Riedel-Kruse et al., 2011; Gerber et al., 2016a), and fungi (Kim et al., 2019) have been incorporated, and others such as viruses (Kim, 2021) have been proposed. **Stimulus modalities** included chemicals

(Riedel-Kruse et al., 2011; Hossain et al., 2015), light (Lee et al., 2015a; Hossain et al., 2016), and electric fields (Bakkum et al., 2007; Riedel-Kruse et al., 2011), and mechanical mechanisms (Ofer et al., 2021). *Euglena gracilis* (Figures 1B,C) has been proven to be particularly amenable for real-time HBI as it is easy to purchase and culture long term. Furthermore, it has a very robust directional negative phototaxis movement (swimming speed of  $\sim 50 \mu\text{m/s}$ , reaction time of  $\sim 0.5 \text{ s}$ ) as well as many other interesting responses to various light stimuli (Gerber et al., 2016a; Lam et al., 2017; Tsang et al., 2018), its comparably large (length of  $\sim 50 \mu\text{m}$ ) and colorful characteristics facilitate an easy and pleasant microscopy experience, and it is already in wide educational use (Oswald and Kwiatkowski, 2011). When choosing biological materials to work with, a number of considerations should be followed (Gerber et al., 2016b), especially regarding the desired response type, response time and response robustness for the desired interactivity; the particular application, e.g., education or art, also further defines these choices. As a stimulus, light is particularly advantageous given its speed and ease of control (Lam et al., 2017). While genetically engineered or "synthetic organisms" have not yet been appreciably deployed toward HBI (Huang et al., 2018), they hold promise for tailoring desired interactions even further.

The biological material needs to be housed, maintained, stimulated, and observed ("**hardware**," Figure 2B). Therefore, it is necessary to distinguish between hands-on and one-time only activities with logistics similar to classic educational experiments (Gerber et al., 2016a; Kim et al., 2016) and long-term robust activities spanning over days to months that subsequently require a significant degree of robustness and automation (Lee et al., 2015a; Hossain et al., 2016; Lam et al., 2019). In particular, **microfluidic technology** (Merrin, 2019; Rackus et al., 2019) with various state measurements on the quality of the biological material, feedback to correct for the desired behavior, and parallelization of multiple systems have been proven to be effective over weeks in a demonstrated **real-time interactive biology cloud lab** (Hossain et al., 2016; Hossain and Riedel-Kruse, 2018). Nevertheless, further improvements are desirable. Overall, advances in open source electronics, such as Arduino and Raspberry Pi and simple webcam microscopes and associated optics provide low-cost and effective technologies for setting up such systems (Hossain et al., 2016; Kim et al., 2016). Microfluidic technologies are not yet as accessible and user-friendly as DIY digital and electronic technologies, although they are steadily improving (Gerber et al., 2015; Rackus et al., 2019). More complex spatio-temporal motion control of cells has been achieved through structured microfluidic chips (such as a Pac-Man maze) and spatial light fields (Lam et al., 2017, 2020). Robotics technology based on LEGO or DIY systems is also enabling (Hossain et al., 2015; Gerber et al., 2017; Gome et al., 2019; Fuhrmann et al., 2021); i.e., turn-based biology cloud labs with activities spanning days have been demonstrated utilizing such robots (Hossain et al., 2015). Ultimately, it would be desirable to establish **biotic processing units (BPU)** (Figure 1C) that house microbiological material and enable



its stimulation and observation through well-defined, digital interaction channels — in close analogy to electronic central processing units (CPUs) or graphics processing units (GPUs) (Gerber et al., 2016b; Lam et al., 2017; Washington et al., 2019).

Many interface modalities for human users have been demonstrated (“**user interface**,” **Figure 2C**). These modalities include touch screens (Lee et al., 2015a,b; Lam et al., 2019), smartphones (Kim et al., 2016), joysticks (Riedel-Kruse et al., 2011; Kim et al., 2016), full-body and motion detection (Lam et al., 2019; Lee et al., 2020), and web-interfaces (Hossain et al., 2015, 2016). Accessibility to different user groups has been achieved through museum installations (Lee et al., 2015a), cloud labs (Hossain et al., 2016), and hands-on DIY approaches (Kim et al., 2016). In these various forms of interaction, the system may guide human users to perform designed activities under predefined rules and purposes (e.g., games, experiments), or the system can also be configured to mediate free-form interactions and open play where users may need to devise their own ways to experience the system (Lee et al., 2015a). Regarding human senses, most interactions utilize natural human visual perceptions, but others have been explored such as smell (Riedel-Kruse et al., 2011; Huang et al., 2018) or sound (indirectly through sonification) (Riedel-Kruse et al., 2011). Different interaction time scales were demonstrated, ranging from real-time interactions (Lee et al., 2015a; Hossain et al., 2016; Kim et al., 2016) to longer timescales, such as the slow biotic game “Mould Rush” (Kim et al., 2018b) and turn-based cloud labs (Hossain et al., 2015).

## USER STUDIES, DESIGN INSIGHTS, AND BEST PRACTICES

A growing list of specific **interaction design rules** have been established (Gerber et al., 2016b; Kim et al., 2020) (**Figure 2D**). Some of the micro-HBI systems have already been paired with user studies, some even in a controlled fashion. Most of these studies focused on formal and informal education (Kuznetsov et al., 2013; Lee et al., 2015a; Hossain et al., 2016, 2017b; Lam et al., 2019). A few themes and corresponding evidence emerged from these studies: Attitude changes toward science, improved access to biology experimentation for poorer and underprivileged communities through cloud labs, accessibility even for small children due to touchscreens or full-body experiences, and stimulation of prolonged engagement and interactions among multiple users. Users expressed agency and playfulness (“I liked playing around with the online microscope”), and valued the real biology interactions compared to simulations (and the synergistic combination of both) (“using a real microscope is more exciting than using a computer simulation”). A more dedicated comparison museum study (Lam et al., 2019) investigated different interaction modalities with microbial biology, and reported a positive interest in these biological cells due to direct microbial interactions via full-body experiences or touchscreens, whereas the more traditional approach of only controlling

the microscope with a joystick even led to loss of interest for some users.

A unique aspect of HBI is the variability (randomness, noise) that exist in the biological matter. Due to this biological variability, not all cells respond identically to the same stimuli, and the responses also vary over time, which can be viewed as glitches in the system but can also serve to entertain and to further convey that this is a real biological system and not a simulation (Kim et al., 2019; Lee et al., 2020). In designing user activities with clear goals (such as a game), striking a balance to control these variations to enable a robust interaction experience while also utilizing them to convey a sense of realness and authenticity to the biological processes is necessary.

HBI also raises potential ethical questions to the designers and the users. These concerns have significant conceptual overlap with other bioethical discussions (Harvey et al., 2014), such as the ethical principles established in the domain of plant/animal interactions (Mancini, 2011; Aspling et al., 2016) and the ethics discussions in the BioArt domain (Stracey, 2009). Related arguments have also been made regarding a non-anthropocentric design approach for human-plant interactions (Fell et al., 2020). The raised concerns often also stem from a misunderstanding of the microscopic biological subject matter (Harvey et al., 2014). Given the non-sentient nature of these microbiological systems, many of these concerns are usually straightforward to address. Nevertheless, these concerns should be taken seriously when designing such systems to engage the public in a positive and supportive manner (Harvey et al., 2014; Merritt et al., 2020; Pataranutaporn et al., 2020; Gough et al., 2021).

Overall, these user studies as well as various exploratory projects revealed a number of prominent features compared to existing modalities to engage with microbiology; in particular, these user studies subsequently revealed a growing set of **design rules and general design principles** that should be considered when macroscopic humans interact with microscopic living matter (Gerber et al., 2016b; Fell et al., 2020; Merritt et al., 2020; Pataranutaporn et al., 2020; Gough et al., 2021). In addition to more general HCI design principles (Shneiderman, 2016), the following warrant consideration: (i) Biological behavior and responses need to be sufficiently robust, (ii) the interactive time and length scales between microscopic biology and macroscopic humans need to be matched (**Figure 2D**), (iii) the biological variability should be managed but also embraced as a feature, (iv) the activities should convey that this is real biology and not a simulation, (v) potential safety and ethical aspects need to be addressed, and (iv) the specific audience and application needs to be considered.

## DISCUSSION AND OPEN QUESTIONS

The field of micro-HBI is now well over a decade old, with the number of contributors from different fields steadily increasing. Significant advancements have been made regarding technology and interaction design. Furthermore, a large application and design space has been explored (**Figure 2**). These detailed developments are driven for multiple reasons; i.e., technology is

advancing (such as DIY microfluidics, electronics and optics), the field has achieved a large number of demonstration and use cases to highlight its versatility, creative potential and overall feasibility have been demonstrated, and the need for society-wide formal and informal education on these subject matters is established. We suggest that the field should now move beyond conceptual papers and speculative design and focus on more practical advancements and real world applications. In particular, recommendations include enabling technology that is practical, robust, and accessible, and that enables versatile design; endeavors aiming to solve practical needs, such as large-scale education, while also including future business models; and actual user studies that are performed to drive design principles. We conclude key focus points that would drive the field forward:

1. Can we have HBI systems that fully automate long-term and robust interactivity with microbes? (Hossain et al., 2016)
2. Can we safely and ethically use synthetic or genetically modified organisms? (Stark et al., 2018)
3. How can we manufacture, deploy and even personalize HBI systems at scale? (Hossain et al., 2017b)
4. Can HBI have the same social and economic impact as predecessor technologies such as electronic video games? (Gerber et al., 2016b)
5. Can HBI make contributions for citizen science? (Lee et al., 2014; Das et al., 2019)
6. Can HBI implement more complex genres of inorganic counterparts, such as real-time strategy games? (Das et al., 2019)
7. Can we have a commercially available (general purpose) platform (“BPU”) that is accessible and upon which others can design, e.g., a mini-game or Tamagotchi-like system? (Gerber et al., 2016b)

## REFERENCES

- Alistar, M., and Pever, M. (2020). “Semina aeternitatis: using bacteria for tangible interaction with data,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, (Honolulu, HI), 1–13.
- Aspling, F., Wang, J., and Juhlin, O. (2016). “Plant-computer interaction, beauty and dissemination,” in *Proceedings of the Third International Conference on Animal-Computer Interaction*, (Milton Keynes), 1–10.
- Bakkum, D. J., Gamblen, P. M., Ben-Ary, G., Chao, Z. C., and Potter, S. M. (2007). Meart: the semi-living artist. *Front. Neurobot.* 1, 2007. doi: 10.3389/neuro.12.005.2007
- Chen, D., Seong, Y. A., Ogura, H., Mitani, Y., and Sekiya, N. (2021). “Nukabot: design of care for human-microbe relationships,” in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, (Yokohama), 1–7.
- Cheok, A. D., Kok, R. T., Tan, C., Newton Fernando, O. N., Merritt, T., and Sen, J. Y. P. (2008). “Empathetic living media,” in *Proceedings of the 7th ACM Conference on Designing Interactive Systems*, (Cape Town: ACM), 465–473.
- Das, R., Keep, B., Washington, P., and Riedel-Kruse, I. H. (2019). Scientific discovery games for biomedical research. *Ann. Rev. Biomed. Data Sci.* 2, 253–279. doi: 10.1146/annurev-biodatasci-072018-021139
- Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., et al. (2012). Increased diels-alderase activity through backbone remodeling guided by foldit players. *Nat. Biotechnol.* 30, 190–192. doi: 10.1038/nbt.2109
- Fell, J., Greene, T., Wang, J.-C., and Kuo, P.-Y. (2020). “Beyond human-centered design: proposing a biocentric view on design research involving vegetal

8. How can we more effectively bring different fields of expertise together, especially bioengineers, microbiologists, human-computer interaction designers, as well as stakeholders such as educators or the gaming industry? Should there be a dedicated conference on this subject matter?

In conclusion, the argument has been made that micro-HBI might undergo a similar exponential path as its computational counterparts (Gerber et al., 2016b). However, whether that will happen likely significantly depends on standardized and easy-to-use technology of sufficient robustness (e.g., BPU-**Figure 1**) that would open a large design space, enable killer applications (such as medical diagnostics), and support the engagement and curiosity of scientists/engineers, artists, and the general public.

## AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## FUNDING

Funding was provided by the College of Science and the MCB Department at the University of Arizona (IR-K), and by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (no. 2021R1C1C101290011) (SL).

## ACKNOWLEDGMENTS

We are grateful to A. Hamby and G. Day for feedback and discussion on earlier versions of the manuscript.

- subjects,” in *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, (Eindhoven), 209–214.
- Fuhrmann, T., Ahmed, D. I., Arikson, L., Wirth, M., Miller, M. L., Li, E., et al. (2021). “Scientific inquiry in middle schools by combining computational thinking, wet lab experiments, and liquid handling robots,” in *Interaction Design and Children*, (Athens), 444–449.
- Gerber, L. C., Calasanz-Kaiser, A., Hyman, L., Voitiuk, K., Patil, U., and Riedel-Kruse, I. H. (2017). Liquid-handling lego robots and experiments for stem education and research. *PLoS Biol.* 15, e2001413. doi: 10.1371/journal.pbio.2001413
- Gerber, L. C., Doshi, M. C., Kim, H., and Riedel-Kruse, I. H. (2016a). “Biographr: Science games on a biotic computer,” in *Proceedings of the First Joint International Conference on Digital Games Research Association and Foundation of Digital Games DiGRA/FDG* (Dundee).
- Gerber, L. C., Kim, H., and Riedel-Kruse, I. H. (2015). Microfluidic assembly kit based on laser-cut building blocks for education and fast prototyping. *Biomicrofluidics* 9, 064105. doi: 10.1063/1.4935593
- Gerber, L. C., Kim, H., and Riedel-Kruse, I. H. (2016b). “Interactive biotechnology: design rules for integrating biological matter into digital games,” in *DiGRA/FDG* (Dundee).
- Glass, D. S., and Riedel-Kruse, I. H. (2018). A synthetic bacterial cell-cell adhesion toolbox for programming multicellular morphologies and patterns. *Cell* 174, 649–658. doi: 10.1016/j.cell.2018.06.041
- Gome, G., Waksberg, J., Grishko, A., Wald, I. Y., and Zuckerman, O. (2019). “Openlh: open liquid-handling system for creative experimentation with biology,” in *Proceedings of the Thirteenth International Conference on Tangible*,

- Embedded, and Embodied Interaction, TEI '19* (New York, NY: Association for Computing Machinery), 55–64.
- Gough, P., Yoo, S., Tomitsch, M., and Ahmadpour, N. (2021). Applying bioaffordances through an inquiry-based model: a literature review of interactive biodesign. *Int. J. Hum. Comput. Interact.* 37, 1583–97. doi: 10.1080/10447318.2021.1898846
- Harvey, H., Havard, M., Magnus, D., Cho, M. K., and Riedel-Kruse, I. H. (2014). Innocent fun or “microslavery”? an ethical analysis of biotic games. *Hastings Center Rep.* 44, 38–46. doi: 10.1002/hast.386
- Hossain, Z., Bumbacher, E., Blikstein, P., and Riedel-Kruse, I. (2017a). “Authentic science inquiry learning at scale enabled by an interactive biology cloud experimentation lab,” in *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, (Cambridge, MA: ACM), 237–240.
- Hossain, Z., Bumbacher, E., Brauneis, A., Diaz, M., Saltarelli, A., Blikstein, P., et al. (2017b). Design guidelines and empirical case study for scaling authentic inquiry-based science learning via open online courses and interactive biology cloud labs. *Int. J. Artif. Intell. Educ.* 28, 478–507. doi: 10.1007/s40593-017-0150-3
- Hossain, Z., Bumbacher, E. W., Chung, A. M., Kim, H., Litton, C., Walter, A. D., et al. (2016). Interactive and scalable biology cloud experimentation for scientific inquiry and education. *Nat. Biotechnol.* 34, 1293–1298. doi: 10.1038/nbt.3747
- Hossain, Z., Jin, X., Bumbacher, E. W., Chung, A. M., Koo, S., Shapiro, J. D., et al. (2015). “Interactive cloud experimentation for biology: An online education case study,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul: ACM), 3681–3690.
- Hossain, Z., Riedel-Kruse, I. H. (2018). “Life-science experiments online: technological frameworks and educational use cases,” in *Cyber-Physical Laboratories in Engineering and Science Education*, eds M. Auer, A. Azad, A. Edwards, and T. de Jong (Cham: Springer). doi: 10.1007/978-3-319-76935-6\_11
- Huang, A., Nguyen, P. Q., Stark, J. C., Takahashi, M. K., Donghia, N., Ferrante, T., et al. (2018). Biobits explorer: a modular synthetic biology education kit. *Sci. Adv.* 4, eaat5105. doi: 10.1126/sciadv.aat5105
- Jin, X., and Riedel-Kruse, I. H. (2018). Biofilm lithography enables high-resolution cell patterning via optogenetic adhesion expression. *Proc. Natl. Acad. Sci. U.S.A.* 115, 3698–3703. doi: 10.1073/pnas.1720676115
- Kim, H., Gerber, L. C., Chiu, D., Lee, S. A., Cira, N. J., Xia, S. Y., et al. (2016). Luduscope: accessible interactive smartphone microscopy for life-science education. *PLoS ONE* 11, e0162602. doi: 10.1371/journal.pone.0168053
- Kim, R. (2021). “Virus as quasi-living bio-material for interaction design: Practical, ethical, and philosophical implications,” in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.
- Kim, R., Thomas, S., Dierendonck, R. V., Kaniadakis, A., and Poslad, S. (2018a). “Microbial integration on player experience of hybrid bio-digital games,” in *International Conference on Intelligent Technologies for Interactive Entertainment* (Yokohama: Springer), 148–159.
- Kim, R., Thomas, S., van Dierendonck, R., Bryan-Kinns, N., and Poslad, S. (2020). “Working with nature’s lag: Initial design lessons for slow biotic games,” in *International Conference on the Foundations of Digital Games*, (Bugibba), 1–4.
- Kim, R., Thomas, S., van Dierendonck, R., and Poslad, S. (2018b). “A new mould rush: designing for a slow bio-digital game driven by living micro-organisms,” in *Proceedings of the 13th International Conference on the Foundations of Digital Games* (Malmo), 1–9.
- Kim, R., van Dierendonck, R., and Poslad, S. (2019). “Moldy ghosts and yeast invasions: glitches in hybrid bio-digital games,” in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, (Glasgow), 1–6.
- Koepnick, B., Flatten, J., Husain, T., Ford, A., Silva, D.-A., Bick, M. J., et al. (2019). De novo protein design by citizen scientists. *Nature* 570, 390–394. doi: 10.1038/s41586-019-1274-4
- Kuznetsov, S., Barrett, C., Fernando, P., and Fowler, K. (2018). “Antibiotic-responsive bioart: exploring diybio as a design studio practice,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC), 1–14.
- Kuznetsov, S., Harrigan-Anderson, W., Faste, H., Hudson, S. E., and Paulos, E. (2013). “Community engagements with living sensing systems,” in *Proceedings of the 9th ACM Conference on Creativity and Cognition, Candamp/C '13* (New York, NY: Association for Computing Machinery), 213–222.
- Kuznetsov, S., Taylor, A. S., Regan, T., Villar, N., and Paulos, E. (2012). “At the seams: diybio and opportunities for hci,” in *DIS '12* (New York, NY: Association for Computing Machinery), 258–267.
- Lam, A. T., Griffin, J., Loeun, M. A., Cira, N. J., Lee, S. A., and Riedel-Kruse, I. H. (2020). “Pac-euglena: a living cellular pac-man meets virtual ghosts,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI), 1–13.
- Lam, A. T., Ma, J., Barr, C., Lee, S. A., White, A. K., Yu, K., et al. (2019). First-hand, immersive full-body experiences with living cells through interactive museum exhibits. *Nat. Biotechnol.* 37, 1238–1241. doi: 10.1038/s41587-019-0272-2
- Lam, A. T., Samuel-Gama, K. G., Griffin, J., Loeun, M., Gerber, L. C., Hossain, Z., et al. (2017). Device and programming abstractions for spatiotemporal control of active micro-particle swarms. *Lab. Chip.* 17, 1442–1451. doi: 10.1039/C7LC00131B
- Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., et al. (2014). Rna design rules from a massive open laboratory. *Proc. Natl. Acad. Sci. U.S.A.* 111, 2122–2127. doi: 10.1073/pnas.1313039111
- Lee, K., Jung, J., and Lee, S. A. (2020). “Microaquarium: an immersive and interactive installation with living microorganisms,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI EA '20* (New York, NY: Association for Computing Machinery), 1–4.
- Lee, S. A., Bumbacher, E., Chung, A. M., Cira, N., Walker, B., Park, J. Y., et al. (2015a). “Trap it!: a playful human-biology interaction for a museum installation,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, (New York, NY: ACM), 2593–2602.
- Lee, S. A., Chung, A. M., Cira, N., and Riedel-Kruse, I. H. (2015b). “Tangible interactive microbiology for informal science education,” in *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction* (New York, NY: ACM), 273–280.
- Lockee, B. B. (2021). Online education in the post-covid era. *Nat. Electr.* 4, 5–6. doi: 10.1038/s41928-020-00534-0
- Loparev, A., Westendorf, L., Flemings, M., Cho, J., Littrell, R., Scholze, A., et al. (2017). “Bacpack: exploring the role of tangibles in a museum exhibit for bio-design,” in *Proceedings of the Eleventh International Conference on Tangible, Embedded, and Embodied Interaction, TEI '17* (New York, NY: Association for Computing Machinery), 111–120.
- Mancini, C. (2011). Animal-computer interaction: a manifesto. *Interactions* 18, 69–73. doi: 10.1145/1978822.1978836
- Merrin, J. (2019). Frontiers in microfluidics, a teaching resource review. *Bioengineering* 6, 109. doi: 10.3390/bioengineering6040109
- Merritt, T., Hamidi, F., Alistar, M., and DeMenezes, M. (2020). Living media interfaces: a multi-perspective analysis of biological materials for interaction. *Digital Creativity* 31, 1–21. doi: 10.1080/14626268.2019.1707231
- Nguyen, P. Q., Soenksen, L. R., Donghia, N. M., Angenent-Mari, N. M., de Puig, H., Huang, A., et al. (2021). Wearable materials with embedded synthetic biology sensors for biomolecule detection. *Nat. Biotechnol.* 39, 1366–1374. doi: 10.1038/s41587-021-00950-3
- Ofer, N., Bell, F., and Alistar, M. (2021). Designing direct interactions with bioluminescent algae. *Design. Interact. Syst. Conf.* 2021, 1230–1241. doi: 10.1145/3461778.3462090
- Osthoft, S. (2001). *Eduardo Kac’s Genesis: Biotechnology Between the Verbal, the Visual, the Auditory, and the Tactile*. Cambridge, MA: MIT Press.
- Oswald, C., and Kwiatkowski, S. (2011). Population growth in euglena: a student-designed investigation combining ecology, cell biology, and quantitative analysis. *Am. Biol. Teach.* 73, 469–473. doi: 10.1525/abt.2011.73.8.8
- Pataranutaporn, P., and Lyle, K. (2018). “Toward human-magic interaction: interfacing biological, tangible, and cultural technology,” in *HCI International 2018 - Posters’ Extended Abstracts*, ed C. Stephanidis (HCI International).
- Pataranutaporn, P., Vujic, A., Kong, D. S., Maes, P., and Sra, M. (2020). “Living bits: opportunities and challenges for integrating living microorganisms in human-computer interaction,” in *Proceedings of the Augmented Humans International Conference* (Kaiserslautern), 1–12.
- Rackus, D. G., Riedel-Kruse, I. H., and Pamme, N. (2019). Learning on a chip: microfluidics for formal and informal science education. *Biomicrofluidics* 13, 041501. doi: 10.1063/1.5096030

- Riedel-Kruse, I. H., Chung, A. M., Dura, B., Hamilton, A. L., and Lee, B. C. (2011). Design, engineering and utility of biotic games. *Lab. Chip* 11, 14–22. doi: 10.1039/C0LC00399A
- Shneiderman, B., Plaisant, C., Cohen, M. S., Jacobs, S., Elmqvist, N., and Diakopoulos, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 6th Edn. Available online at: [https://nsworks.nova.edu/gscis\\_facbooks/18](https://nsworks.nova.edu/gscis_facbooks/18)
- Smith, R. S. H., Bader, C., Sharma, S., Kolb, D., Tang, T.-C., Hosny, A., et al. (2020). Hybrid living materials: digital design and fabrication of 3d multimaterial structures with programmable biohybrid surfaces. *Adv. Funct. Mater.* 30, 1907401. doi: 10.1002/adfm.201907401
- Stark, J. C., Huang, A., Nguyen, P. Q., Dubner, R. S., Hsu, K. J., Ferrante, T. C., et al. (2018). Biobits bright: a fluorescent synthetic biology education kit. *Sci. Adv.* 4, eaat5107. doi: 10.1126/sciadv.aat5107
- Stojanovic, M. N., and Stefanovic, D. (2003). A deoxyribozyme-based molecular automaton. *Nat. Biotechnol.* 21, 1069–1074. doi: 10.1038/nbt862
- Stracey, F. (2009). Bio-art: the ethics behind the aesthetics. *Nat. Rev. Mol. Cell Biol.* 10, 496. doi: 10.1038/nrm2699
- Tsang, A. C., Lam, A. T., and Riedel-Kruse, I. H. (2018). Polygonal motion and adaptable phototaxis via flagellar beat switching in *euglena gracilis*. *bioRxiv* 292896. doi: 10.1101/292896
- van Eck, W., and Lamers, M. H. (2018). “Mapping the field of organism-involved computer games,” in *Proceedings of the 13th International Conference on the Foundations of Digital Games* (Malmo), 1–8.
- Washington, P., Samuel-Gama, K. G., Goyal, S., Ramaswami, A., and Riedel-Kruse, I. H. (2019). Interactive programming paradigm for real-time experimentation with remote living matter. *Proc. Natl. Acad. Sci. U.S.A.* 116, 5411–5419. doi: 10.1073/pnas.1815367116
- Wilson, A. B., Taylor, M. A., Klein, B. A., Sugrue, M. K., Whipple, E. C., and Brokaw, J. J. (2016). Meta-analysis and review of learner performance and preference: virtual versus optical microscopy. *Med. Educ.* 50, 428–440. doi: 10.1111/medu.12944
- Yao, L., Ou, J., Cheng, C.-Y., Steiner, H., Wang, W., Wang, G., et al. (2015). “bioLogic: natto cells as nanoactuators for shape changing interfaces,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul), 1–10.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Lee and Riedel-Kruse. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





## OPEN ACCESS

## EDITED BY

Jonni Virtema,  
The University of Sheffield,  
United Kingdom

## REVIEWED BY

Shaul Almagor,  
Technion Israel Institute of  
Technology, Israel  
Engel Lefauchaux,  
Inria Nancy-Grand-Est Research  
Centre, France

## \*CORRESPONDENCE

Daniele Dell'Erba  
daniele.dell-erba@liverpool.ac.uk  
Sven Schewe  
sven.schewe@liverpool.ac.uk

## SPECIALTY SECTION

This article was submitted to  
Theoretical Computer Science,  
a section of the journal  
Frontiers in Computer Science

RECEIVED 05 May 2022

ACCEPTED 26 August 2022

PUBLISHED 20 September 2022

## CITATION

Dell'Erba D and Schewe S (2022)  
Smaller progress measures and  
separating automata for parity games.  
*Front. Comput. Sci.* 4:936903.  
doi: 10.3389/fcomp.2022.936903

## COPYRIGHT

© 2022 Dell'Erba and Schewe. This is  
an open-access article distributed  
under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#).  
The use, distribution or reproduction  
in other forums is permitted, provided  
the original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which  
does not comply with these terms.

# Smaller progress measures and separating automata for parity games

Daniele Dell'Erba\* and Sven Schewe\*

Department of Computer Science, University of Liverpool, Liverpool, United Kingdom

Calude et al. have recently shown that parity games can be solved in quasi-polynomial time, a landmark result that has led to several approaches with quasi-polynomial complexity. Jurdzinski and Lazic have further improved the precise complexity of parity games, especially when the number of priorities is low (logarithmic in the number of positions). Both of these algorithms belong to a class of game solving techniques now often called separating automata: deterministic automata that can be used as witness automata to decide the winner in parity games up to a given number of states and colors. We suggest several adjustments to the approach of Calude et al. that lead to smaller statespaces. These include and improve over those earlier introduced by Fearnley et al. We identify two of them that, together, lead to a statespace of exactly the same size Jurdzinski and Lazic's concise progress measures, which currently hold the crown as the smallest statespace. The remaining improvements, hence, lead to a further reduction in the size of the statespace, making our approach the most succinct progress measure available for parity games.

## KEYWORDS

parity games, progress measures, value iteration, separating automata, quasi-polynomial algorithms

## 1. Introduction

Parity games are two-player perfect information turn-based zero-sum games of infinite duration played on finite directed graphs. Each vertex, that is labeled with an integer *color*, is assigned to one of the two players *even* or *odd*. A play consists of an infinite sequence of players' moves around the graph, and the winner is determined by the *parity* of the largest color encountered along the play. Hence, the player even wins if it is an even color, and player odd wins otherwise.

Parity games have been extensively studied for their practical applications, to determine their complexity status, and to find efficient solutions.

From a practical point of view, many problems in formal verification and synthesis can be reformulated in terms of solving parity games. Computing winning strategies for these games is linear-time equivalent to solving the modal  $\mu$ -calculus model-checking problem (Emerson and Lei, 1986; Emerson et al., 2001). Parity games can be applied to solve the complementation problem for alternating automata (Grädel et al., 2002) or the

emptiness of the corresponding nondeterministic tree automata (Kupferman and Vardi, 1998). These automata, in turn, can be used to solve the satisfiability and model-checking problems for several expressive logics (Berwanger and Grädel, 2004; Chatterjee et al., 2010; Mogavero et al., 2010, 2012; Benerecetti et al., 2013), such as  $\mu$ -calculus (Wilke, 2001; Schewe and Finkbeiner, 2006) and ATL\* (Alur et al., 2002; Schewe, 2008).

On the complexity-theoretic side, determining the winner of a parity game is a problem that lies in  $\text{NP} \cap \text{co-NP}$  (Emerson et al., 2001), being memoryless determined (Martin, 1975; Emerson and Jutla, 1991; Mostowski, 1991), and it has even been proved to belong to  $\text{UP} \cap \text{co-UP}$  (Jurdziński, 1998), and later to be solvable in quasi-polynomial time (Calude et al., 2017). However, determining whether they belong also to  $\text{P}$  is still an open problem.

The existing algorithms for solving parity games can be divided into two classes. The first one collects approaches that solve the game by creating a winning strategy for one of the two players in the entire game. This can be done by either employing a value iteration over progress measures (Jurdziński, 2000) or iteratively improving the current strategy (Vöge and Jurdziński, 2000; Fearnley, 2010; Friedmann, 2013). To the second class belong approaches that decompose the solution of a game into the analysis of its subgames in a divide-and-conquer concept. To do so, these approaches partition the game into a set of positions that satisfy certain properties. Then, these portions of the game are recursively solved and by suitably composing them also the initial game is solved. The name of the sets employed by the technique depends on the properties: attraction set (Zielonka, 1998; Schewe, 2007), region (Benerecetti et al., 2016, 2018a,b), tangle (van Dijk, 2018), and justification (Lapauw et al., 2020).

Many algorithms from both classes have been refined to achieve a quasi-polynomial upper bound since the contribution of Calude et al. (2017). This seminal algorithm works as a value-iteration approach with compact measures for which a poly-logarithmic size witness is sufficient, rather than storing the entire history of the play that is exponential. The same approach has been refined improving the complexity result (Fearnley et al., 2017; Jurdziński and Lazic, 2017), while the same complexity has been achieved by several different approaches such as the register-index algorithm (Lehtinen, 2018) and the bounded version of the recursive algorithm (Parys, 2019). Interestingly, all known quasi-polynomial algorithms can be derived by the separation approach which also provides a lower bound for these techniques (Czerwinski et al., 2018).

## 1.1. Contribution

We adjust the definitions of *witnesses*, the data structure first used by Calude et al. (2017) and the way they are updated in several ways. In the following, we provide a high-level idea of

our contribution in comparison with other approaches (Calude et al., 2017; Jurdziński and Lazic, 2017; Fearnley et al., 2019). This description makes use of the notion of witness and can be skipped by non expert readers.

The most clear-cut improvement is the increased succinctness of the resulting structures. The main contributor to this improvement is the one that skips the double occurrence of odd colors in the classic witnesses. Where the highest color is even, this improvement alone obtains a statespace of quite different structure to, but the same size as, the currently smallest statespace from Jurdziński and Lazic approach (Jurdziński and Lazic, 2017). Integrating this with the small refinement from Fearnley et al. (2019), in which when the maximal color is odd, we can just reset the witness to its initial value instead of recording this maximal odd value, extends the refinement to the case where the maximal color is odd.

The other advantages are minor. They include not recording odd colors in the least relevant position (known from Fearnley et al., 2019 broadly halving the statespace) and a novel way of counting the length of even chains represented by witnesses, which broadly halves the statespace at the points where the complexity jumps—when at the even chains that need to be considered reach a new power of 2—but successively loses this advantages, losing it completely which the even chains that need to be represented can reach the next power of 2 minus 1). Integrating these two improvements provide a statespace reduction from just under 2 to just under 4 of our approach over (Jurdziński and Lazic, 2017). A minor additional saving is reached by also skipping minimal colors when they are odd (Fearnley et al., 2019).

The second improvement is a re-definition of the semantics of witnesses, moving from the classic witnesses to *color witnesses*, where all positions with the same color in a witness refer to one chain, instead of referring to many. This re-definition accelerates the convergence, though the acceleration is muted where it is used for value iteration.

## 1.2. Outline

We discuss a variation of the algorithm of Calude et al. (2017), partly in the original version and partly in the variation suggested by Fearnley et al. (2019), to extend this approach to value iteration.

After the general preliminaries, we recap this approach in Section 3, using a mild variation of the witness from Calude et al. (2017) for a basic update rule  $\text{up}'$  that updates a witness  $\mathbf{b}$  when reading a state with color  $v$ , and an antagonistic update rule  $\text{au}'$  that updates a witness  $\mathbf{b}$  when reading a state with color  $v$ .

We then amend those rules in two steps. The first step (Section 4) reduces the statespace but otherwise retains the classic lines of Fearnley et al. (2019). It is a simple extension that carries the easiest to spot improvement of this study: the

reductions from the statespace of the witnesses used, leading to more concise witnesses.

The backbone of the statespace reduction is to simply restrict the number of times an odd color occurs in a witness to at most once. Where the maximal color is even, this change alone leads to a perfect match in size with the statespace of Jurdziński and Lazic (2017), which is currently the smallest. This perfect match is a bit surprising, as the structure of the statespace is very different.

The remaining changes extend the restriction to the case where the maximal color is odd and collect some further minor reductions that roughly lead to a spatespace reduction that is usually in a range between 2 and 4, where the advantage is strongest when the number of states with an even number of colors is a power of 2.

Subsequently, we re-interpret the semantics of a witness in Section 5. This change in semantics does not change the statespace, but it allows for updating the witnesses faster. Faster updating is mainly improving the basic update rule, but to some extent also the antagonistic update rule, leading to faster convergence in both cases.

We then turn to an estimation of the new statespace in Section 6. For this, we proceed in several steps. We first look at the two classic statespaces, considering the previously most concise one those of Calude et al.'s original QP algorithm (Calude et al., 2017).

We then turn on using those improvements to Calude et al. (2017) that lead to a statespace of size equal to that of Jurdziński and Lazic (2017). This is done in order to be able to show that the two statespaces are of precisely the same size, but also to have a clear understanding of which improvements remain beyond this, and to focus on how they influence the statespace.

We then exemplify how the three statespaces compare in size in Section 7.

## 2. Preliminaries

Parity games are turn-based zero-sum games played between two players, even and odd, over finite graphs. A parity game  $\mathcal{P}$  is a tuple  $(V_e, V_o, E, C, \phi)$ , where

- $(V = V_e \cup V_o, E)$  is a finite directed graph, where the set  $V$  of vertices is partitioned into a set  $V_e$  of vertices controlled by the player *even* and a set  $V_o$  of vertices controlled by player *odd*, and where  $E \subseteq V \times V$  is the set of edges;
- $C \subseteq \mathbb{N} = \{1, 2, 3, \dots\}$  is a finite consecutive set of colors, such that  $C = \{1, 2, \dots, \max\{C\}\}$  or  $C = \{2, 3, \dots, \max\{C\}\}$  holds; and
- $\phi: V \rightarrow C$  is the coloring functions that maps each vertex to a color.

We define  $C^- = C \setminus \{\max\{C\}\}$  if the highest color  $\max\{C\}$  is odd, and  $C^- = C$  if the highest color  $\max\{C\}$  is even. We also require that every vertex has at least one outgoing edge.

Intuitively, a parity game  $\mathcal{P}$  is played between the two players by moving a token along the edges of the directed graph  $(V, E)$ . A play of such a game starts at some initial vertex  $v_0 \in V$  where the token is placed at the beginning. The player controlling this vertex then chooses a successor vertex  $v_1$  such that  $(v_0, v_1) \in E$ , and the token is moved to this successor vertex. In the next turn, the player controlling the vertex  $v_1$  makes his choice by picking a successor vertex  $v_2$  where to move the token, such that  $(v_1, v_2) \in E$ , and so on. In this manner, both players move the token over the arena and, thus, form an infinite play of the game.

Formally, a play of a game  $\mathcal{P}$  is an infinite sequence of vertices  $\langle v_0, v_1, \dots \rangle \in V^\omega$  such that, for all  $i \geq 0$ , we have that  $(v_i, v_{i+1}) \in E$ . We denote as  $\text{Plays}_{\mathcal{P}}(v)$  the set of plays of the game  $\mathcal{P}$  that origins in a vertex  $v \in V$  and as  $\text{Plays}_{\mathcal{P}}$  the set of all plays of the game. We omit the subscript when the arena is clear from the context. The color mapping  $\phi: V \rightarrow C$  can be extended from vertices to plays by defining the mapping  $\phi: \text{Plays} \rightarrow C^\omega$  as  $\langle v_0, v_1, \dots \rangle \mapsto \langle \phi(v_0), \phi(v_1), \dots \rangle$ .

A play  $\langle v_0, v_1, \dots \rangle$  is won by the player *even* if  $\limsup_{i \rightarrow \infty} \phi(v_i)$  is even, and by player *odd* otherwise.

A *prefix* of a play (or play prefix) is a non-empty initial sequence  $\langle v_0, v_1, \dots, v_m \rangle$  of a play  $\langle v_0, v_1, \dots \rangle$ .

For a play  $\rho = \langle v_0, v_1, \dots \rangle$  or play prefix  $\rho = \langle v_0, v_1, \dots, v_n \rangle$ , an *even chain* of length  $\ell$  is a sequence of positions  $p_1 < p_2 < p_3 < \dots < p_\ell$  (with  $0 \leq p_1$  and, for plays prefixes,  $p_\ell \leq n$ ) in  $\rho$  that has the following properties:

- for all  $j \in \{1, \dots, \ell\}$ , we have that  $\phi(v_{p_j})$  is even, and
- for all  $j \in \{1, \dots, \ell - 1\}$  the colors in the subsequence defined by  $p_j$  and  $p_{j+1}$  are less than or equal to  $\phi(p_j)$  or  $\phi(p_{j+1})$ . More formally, we have that all colors  $\phi(v_{p_j}), \phi(v_{(p_j)+1}), \dots, \phi(v_{p_{(j+1)}})$  are less than or equal to  $\max\{\phi(v_{p_j}), \phi(v_{p_{j+1}})\}$ .

A strategy for the player *even* is a function  $\sigma: V^*V_e \rightarrow V$  such that  $(v, \sigma(\rho, v)) \in E$  for all  $\rho \in V^*$  and  $v \in V_e$ . If a strategy  $\sigma$  only depends on the last state, then is called memoryless ( $\sigma(\rho, v) = \sigma(\rho', v)$  for all  $\rho, \rho' \in V^*$  and  $v \in V_e$ ). A play  $\langle v_0, v_1, \dots \rangle$  is consistent with  $\sigma$  if, for every prefix  $\rho_n = v_0, v_1, \dots, v_n$  of the play that ends in a state of player *even* ( $v_n \in V_e$ ),  $\sigma(\rho_n) = v_{n+1}$  holds. The player *even* wins the game starting at  $v_0$  if the player has a strategy  $\sigma$  such that either all plays  $\langle v_0, v_1, \dots \rangle$  consistent with  $\sigma$  satisfying  $\limsup_{i \rightarrow \infty} \phi(v_i)$  (i.e., the highest color that occurs infinitely often in the play) is even, that, being the game finite (i.e., the number of states is finite), simplifies in all plays  $\langle v_0, v_1, \dots \rangle$  consistent with  $\sigma$  contain a loop  $v_i, v_{i+1}, \dots, v_{i+k}$ , that satisfies  $v_i = v_{i+k}$  and that  $\max\{\phi(v_i), \dots, \phi(v_{i+j})\}$  is even. In both cases,  $\sigma$  might be memoryless.

A *separating automaton* (Bojańczyk and Czerwiński, 2018) for parity games with a set of colors  $C$  and a bounded number of up to  $b$  states with even color<sup>1</sup>, is a deterministic reachability automaton  $\mathcal{A} = (Q, C, q_0, \delta, \text{won})$ , where

- $Q$  is the set of states, with  $q_0, \text{won} \in Q$ ,  $q_0$  is the initial state and  $\text{won}$  is the target state (and sink), and
- $\delta : Q \times C \rightarrow Q$  is the transition function (with  $\delta(\text{won}, v) = \text{won}$  for all  $v \in C$ ),

such that, for all parity games with colors  $\subseteq C$  that have no more than  $b$  states of even color, the following holds:

- if  $v \in V$  is a winning state for the player *even*, then there is a positional strategy  $\sigma$  for the player *even* such that, for every play  $\rho$  from  $v$  consistent with  $\sigma$ ,  $\phi(\rho)$  is accepted by  $\mathcal{A}$  (i.e.,  $\mathcal{A}$  reaches the target state  $\text{won}$ ); and
- if  $v \in V$  is a winning state for player *odd*, then there is a positional strategy  $\sigma$  for player *odd* such that, for every play  $\rho$  from  $v$  consistent with  $\sigma$ ,  $\phi(\rho)$  is rejected by  $\mathcal{A}$  (i.e.,  $\mathcal{A}$  does not reach the target state  $\text{won}$ ).

### 3. Classic Witnesses

We adjust the approach from Calude et al. (2017) and Fearnley et al. (2019), and this section is predominantly taking the representation from Fearnley et al. (2019). It does, however, change some details in the definitions of *i*-witnesses that end in an odd priority and the definition of the value of a witness slightly to suit the rest of the article better. Where the proofs are affected, they are adjusted and given, but the proofs are mostly unaffected by these minor details.

#### 3.1. Classic forward witness

We start with describing the *old* witness without making its semantics formal (as we do not need it in this article), and will turn to the new *concise witness* (Section 4) and the *color witness* (Section 5) afterwards.

##### 3.1.1. *i*-Witnesses

Let  $\rho = v_1, v_2, \dots, v_m$  be a prefix of a play of the parity game. An *even i-witness* is a sequence of (not necessarily consecutive) positions of  $\rho$

$$p_1, p_2, p_3, \dots, p_{2^i},$$

<sup>1</sup> The bound  $b$  is often given instead of the number of states.

of length, exactly  $2^i$ , and an *odd i-witness* is a sequence of (not necessarily consecutive) positions of  $\rho$

$$p_0, p_1, p_2, \dots, p_{2^i}$$

of length exactly  $2^i + 1$ , that satisfy the following properties:

- **Position:** Each  $p_j$  specifies a position in the play  $\rho$ , so each  $p_j$  is an integer that satisfies  $1 \leq p_j \leq m$ .
- **Order:** The positions are ordered. Thus, we have  $p_j < p_{j+1}$  for all  $j < 2^i$ .
- **Evenness:** All positions but the final one are even. Formally, for all  $j < 2^i$ , the color  $\phi(v_{p_j})$  of the vertex in position  $p_j$  is even.  
For position  $p_{2^i}$ , its color  $\phi(v_{p_{2^i}})$  is even for an *even i-witness*, and odd for an *odd i-witness*.
- **Inner domination:** The color of every vertex between  $p_j$  and  $p_{j+1}$  is dominated by the color of  $p_j$  or the color of  $p_{j+1}$ . Formally, for all  $j < 2^i$ , the color of every vertex in the subsequence  $v_{p_j}, v_{p_j+1}, \dots, v_{p_{j+1}}$  is less than or equal to  $\max\{\phi(v_{p_j}), \phi(v_{p_{j+1}})\}$ .
- **Outer domination:** The color of  $p_{2^i}$  is greater than or equal to the color of every vertex that appears after  $p_{2^i}$  in  $\rho$ . Formally, for all  $k$  in the range  $p_{2^i} < k \leq m$ , we have that  $\phi(v_k) \leq \phi(v_{p_{2^i}})$ .

It follows from these properties that an *i-witness* contains an even chain of length  $2^i$ .

##### 3.1.2. Witnesses

We define  $C_- = C^- \cup \{\_ \}$  as the set of colors plus the  $\_$  symbol. A *witness* is a sequence<sup>2</sup>

$$b_k, b_{k-1}, \dots, b_1, b_0,$$

such that each element  $b_i \in C_-$ , and that satisfies the following properties:

- **Witnessing:** there exists a family of *i-witnesses*, one for each element  $b_i$  with  $b_i \neq \_$ . We refer to such an *i-witness* in the run  $\rho$ . We will refer to this witness as

$$p_{i,1}, p_{i,2}, \dots, p_{i,2^i}$$

for even *i-witnesses* and

$$p_{i,0}, p_{i,1}, \dots, p_{i,2^i}$$

<sup>2</sup> While  $k$  can be viewed as "big enough" or as "of arbitrary size" for the definition, we will later see that a length  $k+1$ , with  $k = \lfloor \log_2(e) \rfloor$ , where  $e$  is the number of vertices with an even color, or any other sufficient criterion for the maximal length of an even chain, is sufficient.



for odd  $i$ -witnesses. Thus, even  $i$ -witnesses are in particular even chains of length  $2^i$ , while odd  $i$ -witnesses extended even chains that start with an even chain of length  $2^i$  (and are extended by a further position).

- **Dominating color:** For each  $b_i \neq \_$ , we have that  $b_i = \phi(v_{P_{i,2^i}})$ . That is,  $b_i$  is the outer domination color of the  $i$ -witness.
- **Ordered sequences:** The  $i$ -witness associated with  $b_i$  starts after a  $j$ -witness associated with  $b_j$  whenever  $i < j$ . Formally, for all  $i$  and  $j$  with  $i < j$ , if  $b_i \neq \_$  and  $b_j \neq \_$ , then  $p_{j,2^j} < p_{i,1}$  when the  $i$ -witness is even, and  $p_{j,2^j} < p_{i,0}$  otherwise.

For a little bit of extra conciseness, we also require that  $b_0$  is either even or  $\_$ .

Note that the witness does not store the  $i$ -witnesses associated with each position  $b_i$ . However, the sequence is a witness only if the corresponding  $i$ -witnesses *exist*. Moreover, the colors in a witness are monotonically increasing for growing indices (and thus increase from right to left), since each color  $b_j$  (weakly) dominates all colors that appear afterwards  $\rho$  as a consequence of the dominating color property and the ordered sequences property.

### 3.1.2.1. Forward and backward witnesses

The *forward* witnesses described so far were introduced in Calude et al. (2017), while we now describe the *backward* witnesses and an ordering over them that have been introduced in Fearnley et al. (2019). For each play, prefix  $\rho = v_1, v_2, \dots, v_m$ , we define a reverse play  $\overleftarrow{\rho} = v_m, v_{m-1}, \dots, v_1$ ; a backward witness is a witness for  $\overleftarrow{\rho}$ , or for a prefix of it.

### 3.1.2.2. Order on witnesses

The set  $C_\_$  is ordered by the relation  $\succeq$  such that even numbers are better than odd numbers, higher even numbers are better than smaller even numbers, smaller odd numbers are better than higher odd numbers, and every number is better than  $\_$ . Formally,  $a \succeq b$  if  $b = \_$ ; or  $a$  is even and  $b$  is either odd or  $\_$ ; or  $a \geq b$ ; or  $a \leq b$  and they are both odd.

Using  $\succeq$ , we define an order  $\sqsupseteq'$  over witnesses that compares two witnesses of the same size lexicographically, where the most significant element is  $b_k$  and the least significant element is  $b_0$ . Each element is compared using the order  $\succeq$ . The biggest witness has a special value *won*; i.e.,  $\text{won} \sqsupseteq' \mathbf{b}$  holds for all witnesses  $\mathbf{b}$ .

### 3.1.2.3. The value of a witness

While there is, in principle, a countable set of witnesses, it suffices to take into consideration only those witnesses that do not require the existence of an even chain that is longer than the number states with even color. With this in mind, we define the *value* of a witness as the length of an even chain.

For example,  $\mathbf{b} = 8, 5, 2$  is a forward witness for a run prefix with a color trace 9, 6, 7, 8, 7, 2, 8, 3, 2, 4, 5, 3, 2, 3, 2, where the red, blue, and green color are used to visualize the even

chains (extended, for the sequence in blue). A run for which  $\mathbf{b}$  is a forward witness always includes an even chain of length 6, which consists of the even chain of the leading positions with even priority (in this case: 8) as well as the even chain defined by the first odd color (in this case: 5). For example, this is the even chain shown in cyan: 9, 6, 7, 8, 7, 2, 8, 3, 2, 4, 5, 3, 2, 3, 2.

Formally, we define the following functions for each witness  $\mathbf{b} = b_k, b_{k-1}, \dots, b_0$ :

- **Even positions:**  $\text{even}(\mathbf{b}) = \{i \in \mathbb{N}_0 \mid b_i \text{ is an even number}\}$   
(in the example above, these are positions with index 2 and 0, whose label is 8 and 2, respectively);
- **Relevant  $i$ -witnesses:**  $\text{evenodd}(\mathbf{b}) = \text{even}(\mathbf{b})$  if  $\mathbf{b}$  does not contain an odd number,  
otherwise,  $\text{evenodd}(\mathbf{b}) = \{i \in \text{even}(\mathbf{b}) \mid i > o\} \cup \{o\}$ ,  
with  $o = \max\{i \in \mathbb{N} \mid b_i \text{ is odd}\}$ ;  
(in the example above, these are positions with index 2 and 1, whose label is 8 and 5, respectively); and
- **Value of witness:**  $\text{value}(\mathbf{b}) = \sum_{i \in \text{evenodd}(\mathbf{b})} 2^i$ ;  
(in the example above, this is equal to 6).

**Remark.** The value function from Fearnley et al. (2019) is different from the one defined above, as it uses  $\sum_{i \in \text{even}(\mathbf{b})} 2^i$ . The latter is the sum of the length of the even chains that refer to even entries in the witness.

In the example above, the value function from Fearnley et al. (2019) is the sum of the concatenated even chains in red and green, with a joint length of 5. We will discuss the impact that this difference has on the statespace at the end of Section 6.

We can show that the value of  $\mathbf{b}$  corresponds to the length of an even chain in  $\rho$  that is witnessed by  $\mathbf{b}$ .

**Lemma 1.** Fearnley et al. (2019) If  $\mathbf{b}$  is a (forward or backward) witness of  $\rho$ , then there is an even chain of length  $\text{value}(\mathbf{b})$  in  $\rho$ .

If we count the number of vertices with even colors in the game as  $e = |\{v \in V : \phi(v) \text{ is even}\}|$ , then we can observe that in case we have an even chain longer than  $e$  then  $\rho$  contains a cycle, as there is a vertex with even color visited twice in this even chain. Moreover, the cycle is winning for the player *even*, since the largest priority of its vertices must be even. As a consequence, if the player *even* can force a play that has a witness whose value is strictly greater than  $e$ , *even* wins the game.

**Lemma 2.** Fearnley et al. (2019) If, from an initial state  $v_0$ , the player *even* can force the game to run through a sequence  $\rho$ , such that  $\rho$  has a (forward or backward) witness  $\mathbf{b}$  such that  $\text{value}(\mathbf{b})$  is greater than the number of vertices with even color, then player *even* wins the parity game starting at  $v_0$ .

For this reason, we only need witnesses with value  $\leq e$ , as every witness of value  $> e$  must contain a winning cycle.

We will refer to the set of classic witnesses as  $\mathbb{W} = \{\mathbf{b} \mid \mathbf{b} \text{ is a witness with } \text{value}(\mathbf{b}) \leq e\} \cup \{\text{won}\}$ .

### 3.2. Updating witnesses

Forward witnesses can be constructed incrementally by processing the play one vertex at a time. The following lemmas assume that we have a play prefix  $\rho = v_0, v_1, \dots, v_m$ , and a new vertex  $v_{m+1}$  that we are going to append to  $\rho$  in order to create  $\rho'$ . The value  $d = \phi(v_{m+1})$  denotes the color of the new vertex  $v_{m+1}$ . We will suppose that  $\mathbf{b} = b_k, b_{k-1}, \dots, b_1, b_0$  is a witness for  $\rho$ , and we will construct a witness  $\mathbf{c} = c_k, c_{k-1}, \dots, c_1, c_0$  for  $\rho'$ .

We present three lemmas that allow us to perform this task.

**Lemma 3.** *Fearnley et al. (2019)* Suppose that  $d$  is even, there exists an index  $j$  such that:

- $b_i$  is even for all  $i < j$ ,
- $b_j$  is odd or equal to  $\_$  and
- $b_i \geq d$  or equal to  $\_$  for all  $i > j$ .

If we set  $c_i = b_i$  for all  $i > j$ ,  $c_j = d$ , and  $c_i = \_$  for all  $i < j$ , then  $\mathbf{c}$  is a witness for  $\rho'$ .

Note that we returned to the original definition from Calude et al. (2017) by restricting this updating rule from Lemma 3, called “overflow rule,” to even numbers, whereas the witnesses from Fearnley et al. (2019) also allowed this operation to be performed in the case where  $d$  is odd. The reason for this change is that it reduces the statespace: while this reduction is insignificant in most cases, it is quite substantial if  $e = 2^p - 1$  for some power  $p \in \mathbb{N}$ , as it leads to an increase in the length of the witness. Since statespace reduction is a core target of this article, we opted to be precise here. A full discussion about the statespace is reported in Section 6.

Note that the next lemmas (and their proofs) are essentially independent of the variation of Lemma 3 with respect to the version reported in Fearnley et al. (2019).

**Lemma 4.** *Fearnley et al. (2019)* Suppose that  $d \in C^-$  and there exists an index  $j$  such that:

- $d > b_j \neq \_$  and
- $b_i \geq d$  or equal to  $\_$  for all  $i > j$ .

Then setting  $c_i = b_i$  for all  $i > j$ , setting  $c_j = d$  if  $j \neq 0$  (and  $c_j = \_$  if  $j = 0$ ), and setting  $c_i = \_$  for all  $i < j$  yields a witness for  $\rho'$ .

There is a tiny difference in the proof of this lemma with the one from Fearnley et al. (2019), as we require the length of the  $j$ -witness to be  $2^j + 1$  when  $c_j$  is set to  $d$ . But either  $b_j$  was odd

before, in which case replacing the last index of the  $j$ -witness by  $m + 1$  still produces a witness of length  $2^j + 1$ , or it was even, and in that case, we can instead append  $m + 1$  to the old  $j$ -witness.

**Lemma 5.** *Fearnley et al. (2019)* Suppose that  $d \in C^-$  is odd and, for all  $j \leq k$ , either  $b_j = \_$  or  $b_j \geq d$ . If we set  $c_i = b_i$  for all  $i \leq k$  (i.e., if we set  $\mathbf{c} = \mathbf{b}$ ), then  $\mathbf{c}$  is a witness for  $\rho'$ .

When we want to update a witness with the raw update rule upon scanning another state  $v_{m+1}$  with color  $d = \phi(v_{m+1})$ , we select the according lemma if  $d \in C^-$ . Otherwise, i.e., when  $d = \max\{C\}$  and odd, we re-set the witness to  $\_, \dots, \_$  (which is a witness for every play prefix).

For a given witness  $\mathbf{b}$  and a vertex  $v_{m+1}$ , we denote with

- **Raw update:**  $\text{ru}'(\mathbf{b}, d)$  the raw update of the witness to  $\mathbf{c}$ , as obtained by the update rules described above.
- **Update:**  $\text{up}'(\mathbf{b}, d)$  is either  $\text{ru}'(\mathbf{b}, d)$  if  $\text{value}(\text{ru}'(\mathbf{b}, d)) \leq e$  (where  $e$  is the number of vertices with even color), or  $\text{up}'(\mathbf{b}, d) = \text{won}$ , otherwise.  
In particular,  $\text{up}'(\text{won}, d) = \text{won}$  holds for all  $d \in C$ .
- **Antagonistic update:**  $\text{au}'(\mathbf{b}, d) = \min_{\mathbf{c}' \sqsubseteq' \mathbf{c}} \{\text{up}'(\mathbf{c}', d) \mid \mathbf{b} \sqsubseteq' \mathbf{c} \in \mathbb{W}\}$ .

#### 3.2.1. Basic and antagonistic update game

With these update rules, we define a forward and a backward basic update game played between the two players *even* and *odd*. In this game, they produce a play of the game as usual: if the pebble is in a position assigned to even, then *even* selects a successor, and if the pebble is in a position assigned to odd, then *odd* selects a successor.

Player *even* can stop any time he likes and evaluate the game using  $\mathbf{b}_0 = \_, \dots, \_$  as a starting point and the update rule  $\mathbf{b}_{i+1} = \text{up}'(\mathbf{b}_i, v_i)$  (in the basic update game) and  $\mathbf{b}_{i+1} = \text{au}'(\mathbf{b}_i, v_i)$  (in the antagonistic update game), respectively.

For a forward game, *even* would process the partial play  $\rho^+ = v_0, v_1, v_2, \dots, v_n$  from left to right, and for the backward game he would process the partial play  $\rho^- = v_n, v_{n-1}, \dots, v_0$ . In both cases, *even* has won if, and only if,  $\mathbf{b}_{n+1} = \text{won}$ .

**Theorem 1.** *Fearnley et al. (2019)* Player *even* has a strategy to win the parity game if, and only if, *even* has a strategy to win the classic forward, respectively, backward basic, respectively, antagonistic update game.

This can be formulated in a way that, for a given set  $C$  of colors and  $e$  states with even color, the deterministic reachability automaton with states  $\mathbb{W}$ , initial state  $\_, \dots, \_$ , update rules  $\text{up}'$  (or  $\text{au}'$ ), and reachability goal to reach  $\text{won}$  is a separating automaton.

**Corollary 1.** For a parity game with  $e$  states and  $k = \lfloor \log_2(e) \rfloor$  and  $\mathbb{W}$  the space for witnesses of value  $\leq e$ , length  $k + 1$

and colors  $C$ , both  $\mathcal{U} = (\mathbb{W}; C; \_, \dots, \_, \text{up}'; \text{won})$  and  $\mathcal{A} = (\mathbb{W}; C; \_, \dots, \_, \text{au}'; \text{won})$  are separating automata.  $\square$

The advantage of the antagonistic update rule is that it is monotone:  $\mathbf{b} \sqsubseteq' \mathbf{c} \rightarrow \text{au}'(\mathbf{b}, d) \sqsubseteq \text{au}'(\mathbf{c}, d)$ . This allows for using  $\text{au}'$  in a value iteration algorithm (Fearnley et al., 2019).

## 4. Concise witness

In this article, we suggest a change in the semantics of the witness, and a related reduction of the statespace of witnesses to  $\mathbb{C} \subsetneq \mathbb{W}$ . While this section focuses on condensing the statespace, in the next one we will adjust the semantics and improve the update rule.

The main theoretical advancement is the smaller statespace we will define in this section, as it directly translates into improved bounds, slightly outperforming the currently leading QP algorithm in this regard.

Toward this goal, we define a truncation operator

$$\downarrow_1: \mathbb{W} \rightarrow \mathbb{C}$$

that, for every odd color  $o \in C^-$ , leaves only the leftmost occurrences of  $o$  in a witness and replaces all other occurrences of  $o$  in  $\mathbf{b}$  by  $\_$ .

For example,  $\downarrow_1 \_ , 7, \_ , 7, 5, 4, \_ , 3, 3, \_ , 2 = \_ , 7, \_ , 5, 4, \_ , 3, \_ , 2$ , and  $\downarrow_1 3, 3, 2 = 3, \_ , 2$ . We also have  $\downarrow_1 \text{won} = \text{won}$ .

Note that the definition of  $\downarrow_1$  entails

$$\text{even}(\downarrow_1 \mathbf{b}) = \text{even}(\mathbf{b}),$$

as well as

$$\text{value}(\downarrow_1 \mathbf{b}) = \text{value}(\mathbf{b}).$$

Building on the definition of  $\downarrow_1$ , we continue with the following definitions.

- $\mathbb{C} = \{\downarrow_1 \mathbf{b} \mid \mathbf{b} \in \mathbb{W}\}$ ,
- **Raw update:**  $\text{ru}(\mathbf{b}, d) = \downarrow_1 \text{ru}'(\mathbf{b}, d)$ ,
- **Update:**  $\text{up}(\mathbf{b}, v) = \downarrow_1 \text{up}'(\mathbf{b}, d)$ ,
- **Order over witnesses:** for all  $\mathbf{b}, \mathbf{c} \in \mathbb{C}$ ,  $\mathbf{b} \sqsubseteq \mathbf{c}$  if, and only if,  $\mathbf{b} \sqsubseteq' \mathbf{c}$  (i.e.,  $\sqsubseteq$  is simply a restriction of  $\sqsubseteq'$  from  $\mathbb{W}$  to  $\mathbb{C}$ ), and
- **Antagonistic update:**  $\text{au}(\mathbf{b}, d) = \min_{\sqsubseteq} \{\text{up}(\mathbf{c}, d) \mid \mathbf{b} \sqsupseteq \mathbf{c} \in \mathbb{C}\}$

To justify the use of the concise witness space  $\mathbb{C}$ , we first show that, for all witnesses,  $\mathbf{c} \in \mathbb{W}$  in the *old* witness space, it holds that  $\downarrow_1 \text{ru}'(\mathbf{c}, d) = \downarrow_1 \text{ru}'(\downarrow_1 \mathbf{c}, d)$ . Thus, if we truncate *only* after, or both before *and* after, a raw update does not change the result.

**Lemma 6.** If  $\mathbf{b} = \downarrow_1 \mathbf{c}$ , then  $\text{ru}(\mathbf{b}, d) = \downarrow_1 \text{ru}'(\mathbf{c}, d)$ .

*Proof.* We look at the effect the different update rules have on  $\mathbf{b}$  and  $\mathbf{c}$ . Lemma 3 would (for the same  $j$ ) change the tail (starting with the  $j$ -witness) of  $\mathbf{c}$  and  $\mathbf{b}$  in the same way to  $d, \_, \dots, \_$ , and they either both do or do not satisfy the prerequisites for its application. Thus, the  $\downarrow_1$  operator would remove exactly those positions  $> j$  from  $\text{ru}'(\mathbf{c}, d)$  that it removed from  $\mathbf{c}$ .

When Lemma 4 applies, then it does so for the same index  $j$ , and it simply overrides the tail starting there with  $d, \_, \dots, \_$  (or with  $\_$  if  $j = 0$ ). As all higher positions are unchanged and greater or equal to  $d$ ,  $\mathbf{b} = \downarrow_1 \mathbf{c}$  implies  $\text{ru}(\mathbf{b}, d) = \downarrow_1 \text{ru}'(\mathbf{c}, d)$ .

When the conditions of Lemma 5 apply either for both,  $\mathbf{b}$  and  $\mathbf{c}$  or for neither of them, then we note that Lemma 5 does not change the witness.

Finally, if  $d = \max\{C\}$  and is also odd, then  $\text{ru}'(\mathbf{b}, d) = \text{ru}'(\mathbf{c}, d) = \_, \dots, \_$ , which implies  $\text{ru}(\mathbf{b}, d) = \downarrow_1 \text{ru}'(\mathbf{b}, d) = \downarrow_1 \_, \dots, \_ = \downarrow_1 \text{ru}'(\mathbf{c}, d)$ .  $\square$

This immediately extends to the update rule  $\text{up}/\text{up}'$ , as they differ from  $\text{ru}/\text{ru}'$  only in treating *won*, and to  $\text{au}/\text{au}'$  by monotonicity.

**Corollary 2.** If  $\mathbf{b} = \downarrow_1 \mathbf{c}$ , then  $\text{up}(\mathbf{b}, v) = \downarrow_1 \text{up}'(\mathbf{c}, v)$  and  $\text{au}(\mathbf{b}, v) \sqsupseteq \downarrow_1 \text{au}'(\mathbf{c}, v)$ .  $\square$

The observation that  $\downarrow_1 \text{up}'(\downarrow_1 \mathbf{c}, v) = \text{up}(\downarrow_1 \mathbf{c}, v) = \downarrow_1 \text{up}'(\mathbf{c}, v)$  can be extended to every run prefix: truncating in every step and truncating at the end has the same effect. Thus, by a simple inductive argument, the state *won* is reached with  $\text{up}$  and  $\text{up}'$  at the same time (or not at all in either case).

**Theorem 2.** The player even has a strategy to win the parity game if, and only if, even has a strategy to win the concise forward / backward basic update game.

*Proof.* This follows from Theorem 1: because the same runs are winning when using  $\text{up}$  and  $\text{up}'$  due to Corollary 2, the same player wins the classic and the concise basic update game.  $\square$

**Theorem 3.** The player even has a strategy to win the parity game if, and only if, even has a strategy to win the concise forward / backward antagonistic update game.

*Proof.* For the 'if' case, we observe that Corollary 2 implies with the monotonicity of  $\text{au}$  that, when *even* wins the classic antagonistic update game, *even* also wins the concise antagonistic update game (with the same strategy). Together with Theorem 1, this provides the "if" case.

For the "only if" case, we observe that the monotonicity of  $\text{au}$  entails that, when *odd* wins the basic concise update game, then *odd* wins the antagonistic update game (with the same strategy). Together with Theorem 2, this provides the "only if" case.  $\square$

**Corollary 3.** For a parity game with  $e$  states of even color and  $k = \lfloor \log_2(e) \rfloor$  and  $\mathbb{W}$  the space for witnesses of value  $\leq e$ , length  $k + 1$  and colors  $C$ , both  $\mathcal{U} = (\mathbb{C}; C; \_, \dots, \_, \text{up}; \text{won})$  and  $\mathcal{A} = (\mathbb{C}; C; \_, \dots, \_, \text{au}; \text{won})$  are separating automata.  $\square$

To give intuition to their states, if  $\mathcal{U}$  is in a state  $\mathbf{b} \neq \text{won}$ , it means that  $\mathbf{b}$  is a witness for the play prefix, while  $\text{won}$  means that the play prefix contains an even chain of length  $> e$ , and thus an even cycle.

If  $\mathcal{A}$  is in a state  $\mathbf{b}$  then there is a state  $\mathbf{c} \sqsubseteq \mathbf{b}$  with this property.

As a final remark, in the rare cases where even colors are scarce, their appearance in  $\mathbb{C}$  can also be restricted: if only  $e_\#$  states have an even color  $e$ , then the number of occurrences of  $e$  in a concise witness can be capped to  $e_\#$ , too, as more occurrences of  $e$  without an intermediate occurrence of a higher color would imply that an accepting cycle is in the word.

However, for this to reduce the statespace,  $e_\# \leq \lfloor \log_2(e) \rfloor$  is required, and the closer it comes to  $\lfloor \log_2(e) \rfloor$ , the lesser the saving. In particular, for  $e_\# = \lfloor \log_2(e) \rfloor$ , we would just save a single state.

## 5. Color witnesses

In this section, we use the same data structure as before—the concise witnesses from the previous section—but adjust its semantics.

We introduce two changes to the semantics of witnesses that accelerate the progress to victory for *even* when update operations are made. The changes are applied to the concise witnesses  $\mathbb{C}$ .

Before formalizing how we make our witnesses more flexible and how we use this to re-define the raw update function (and, through this, the update function and the antagonistic update), we describe several examples of how we change the semantics of witnesses.

### 5.1. Motivating examples

#### 5.1.1. Merging witnesses

If we consider the classic witness  $\mathbf{b} = 4, 4, \_$ , it referred to two  $i$ -witnesses that each end on a state with color 4, one of length 4 and one of length 2.

For example,  $\mathbf{b} = 4, 4, \_$  is a classic forward witness for a run prefix with color trace 9, 6, 7, 8, 7, 2, 4, 3, 2, 4 (where the **red** and **blue** color are used to visualize the even chains), whereas 9, 6, 7, 8, 7, 2, 2, 3, 4, 4 is not.

We will instead view this as a *single* color witness for color 4, which then refers to a single even chain of length *at least* 6.

For example,  $\mathbf{b} = 4, 4, \_$  is a forward *color* witness for a run prefix with color trace 9, 6, 7, 8, 7, 2, 2, 3, 4, 4: as we are content with an even chain of length 6, it does not matter that the fourth position of this chain has an entry different to 4.

As a consequence, when passing by a state with the color 6, we can now update the witness to 6, 6, 6, as this would require a single even chain of at least length 7 that ends in a 6.

#### 5.1.2. Shifting witnesses

If we consider the witness  $\mathbf{b} = 4, 2, \_$ , it referred to two  $i$ -witnesses, where the first has length 4 and ends on color 4, while the second has length 2 and ends on color 2.

We will allow making the latter sequence shorter, so long as the former sequence is extended accordingly. For example, when the sequence that ends in 4 has length 5, then it would suffice, for the witness to represent the even chains if the sequence that ends in 2 has length 1.

For example,  $\mathbf{b} = 4, 2, \_$  is a forward color witness for a run prefix with color trace 9, 6, 7, 8, 7, 2, 2, 4, 3, 2 (where the **red** and **blue** colors are used to visualize the even chains)<sup>3</sup>.

Similarly, for  $\mathbf{b} = 6, \_, 4, 2, 2$ , it would be allowed that the length of the even chain that ends in 6 is 18, the subsequent sequence that ends in 4 is 3, and the length of the sequence that ends in 2 is 2. If the length of the sequences ending in 6, 4, and 2 are  $\ell_6$ ,  $\ell_4$ , and  $\ell_2$ , respectively, the constraints would be  $\ell_6 \geq 16$ ,  $\ell_6 + \ell_4 \geq 20$ , and  $\ell_6 + \ell_4 + \ell_2 \geq 23$ .

When passing by a state with the color 8, we can now update the witness to 8, 8,  $\_, \_, \_$ , as this would require a single sequence of at least length 24 that ends in an 8.

#### 5.1.3. Blocked shifting

This shifting cannot be done through an odd color: for  $\mathbf{b} = 4, 3, 2, 2$ , the requirement for the rightmost sequence would be to be of length at least three and to end in a 2. It is, however, possible to shift some of the required lengths of the sequence that ends in 3 to the sequence that ends in 4: if the length of the sequences ending in 4 and 3 are  $\ell_4$  and  $\ell_3$ , respectively, the constraints would be  $\ell_4 \geq 8$  and  $\ell_4 + \ell_3 \geq 13$ . (Recall the odd witnesses need to be one position longer to contain an even chain of the same length.) Thus, reading a 6 would lead to the witness 6, 6,  $\_, 6$ .

## 5.2. Color witness

The biggest change is that an  $i$ -color witness ( $i$ -cowit) refers to the *color*  $i$ , rather than to the position  $b_i$  in the witness. Consequently, we do not have a fixed length of an  $i$ -color witness and refer to the length of the witness for each color  $i$  that appears in a witness as  $\ell_i$ .

As before, we focus in our description on forward witnesses, with backward witnesses being defined accordingly.

<sup>3</sup> Note that  $\mathbf{b} = 4, 2, \_$  is not a classic forward witness for this sequence, whereas  $\mathbf{b}' = 4, \_, 2$  is: the **red even chain** can be shortened (by dropping one 2, e.g., to 9, 6, 7, 8, 7, 2, 2, 4, 3, 2) to an even chain of length 4, which would in and by itself be a 2-witness, whereas the **blue even chain** is merely a 0-witness.



### 5.2.1. $i$ -color witness ( $i$ -cowit) with value $\ell_i$

Let  $\rho = v_1, v_2, \dots, v_m$  be a prefix of a play of the parity game. An *even  $i$ -cowit* is a sequence of (not necessarily consecutive) positions of  $\rho$

$$p_1, p_2, p_3, \dots, p_{\ell_i}$$

of length exactly  $\ell_i$ , and an *odd  $i$ -cowit* is a sequence of (not necessarily consecutive) positions of  $\rho$

$$p_0, p_1, p_2, \dots, p_{\ell_i}$$

of length exactly  $\ell_i + 1$ , that satisfy the following properties:

- **Position:** Each  $p_j$  specifies a position in the play prefix  $\rho$ , so each  $p_j$  is a positive integer that satisfies  $1 \leq p_j \leq m$ .
- **Order:** The positions are ordered. Thus, we have  $p_j < p_{j+1}$  for all  $j < \ell_i$ .
- **Evenness:** All positions but the final one are even. Formally, for all  $j < \ell_i$  the color  $\phi(v_{p_j})$  of the vertex in position  $p_j$  is even.

For position  $p_{\ell_i}$ , its color  $\phi(v_{p_{\ell_i}}) = i$ . Then, the color of that position is even for *even  $i$ -cowit*, and odd for *odd  $i$ -cowit*.

Note that this entails that an  $i$ -cowit has  $\ell_i$  initial even positions that define an even chain of length  $\ell_i$ .

- **Inner domination:** The color of every vertex between  $p_j$  and  $p_{j+1}$  is dominated by the color of  $p_j$  or the color of  $p_{j+1}$ . Formally, for all  $j < \ell_i$ , the color of every vertex in the subsequence  $v_{p_j}, v_{p_j+1}, \dots, v_{p_{j+1}}$  is less than or equal to  $\max\{\phi(v_{p_j}), \phi(v_{p_{j+1}})\}$ .
- **Outer domination:** The color of the vertex  $v_{p_{\ell_i}}$  in position  $p_{\ell_i}$  is  $i$ , i.e.,  $i = \phi(v_{p_{\ell_i}})$ . Moreover,  $i$  is greater than or equal to the color of every vertex that appears after position  $p_{\ell_i}$  in  $\rho$ . Formally, for all  $k$  in the range  $p_{\ell_i} \leq k \leq m$ , we have that  $\phi(v_k) \leq i$ .

### 5.2.2. Color witnesses

Similar to a concise witness, a *color witness* is a sequence

$$b_k, b_{k-1}, \dots, b_1, b_0,$$

of length<sup>4</sup>  $k+1$ , such that each element  $b_i \in C_-$ , and that satisfies the following properties.

- **Properties of the sequence:**

defining  **$i$ -positions** as the positions in  $\mathbf{b}$  that have value  $i$ ,

$\text{positions}(i, \mathbf{b}) = \{j \leq k \mid b_j = i\}$  for every  $i \in C_-$ , the sequence has to satisfy the following constraints:

- **order:** for  $i > j$ , we have that  $b_i \geq b_j$  or  $- \in \{b_i, b_j\}$  holds; and
- **conciseness:** for all odd  $i \in C_-$ ,  $|\text{positions}(i, \mathbf{b})| \leq 1$  and  $b_0 \neq i$  hold.

- **Witnessing:**

- **ordered witnesses:** for  $i > j$  with  $\text{positions}(i, \mathbf{b}) \neq \emptyset$  and  $\text{positions}(j, \mathbf{b}) \neq \emptyset$ , the  $j$ -witness starts after the  $i$ -witness ends. That is  $p_{i, \ell_i} < p_{j, 1}$  if  $j$  is even and  $p_{i, \ell_i} < p_{j, 0}$  if  $j$  is odd.
- using the following definitions,

- **next odd color:**  $\text{odd}(i, \mathbf{b}) = \inf\{j > i \mid j \text{ odd and } \text{positions}(j, \mathbf{b}) \neq \emptyset\}$  defines the next higher odd color than  $i$  that occurs in the color witness (note that  $\text{odd}(i, \mathbf{b}) = \infty$  if no such color exists),

- **unblocked colors:**  $\text{unblocked}(i, \mathbf{b}) = \{j < \text{odd}(i, \mathbf{b}) \mid j \geq i \text{ and } \text{positions}(i, \mathbf{b}) \neq \emptyset\}$  is the set of all colors that are at least  $i$ , but strictly smaller than  $\text{odd}(i, \mathbf{b})$ , and

- **unblocked positions:**  $\text{ubp}(i, \mathbf{b}) = \bigcup_{j \in \text{unblocked}(i, \mathbf{b})} \text{positions}(i, \mathbf{b})$  is the set of positions labeled by an unblocked color,

we have that  $\sum_{\text{unblocked}(i, \mathbf{b})} \ell_i \geq \sum_{j \in \text{ubp}(i, \mathbf{b})} 2^i$  holds

for all  $i \in C_-$ .

It should be noted that neither the  $i$ -cowit-s associated with each color, nor the value of the  $\ell_i$  are stored in a color witness. However, in order for a sequence to be a color witness for an initial sequence of a run, the corresponding  $i$ -cowit-s must exist.

## 5.3. Updating color witnesses

We now show how forward color witnesses can be constructed incrementally by processing the play one vertex at a time. Throughout this subsection, we will suppose that we have a play  $\rho = v_0, v_1, \dots, v_m$ , and a new vertex  $v_{m+1}$  that we would like to append to  $\rho$  to create  $\rho'$ . We will use  $d = \phi(v_{m+1})$  to denote the color of this new vertex. We will suppose that  $\mathbf{b} = b_k, b_{k-1}, \dots, b_1, b_0$  is a color witness for  $\rho$ , and has  $i$ -cowit-s with individual lengths  $\ell_i$ . We will construct a witness  $\mathbf{c} = c_k, c_{k-1}, \dots, c_1, c_0$  for  $\rho'$  and discuss how its inferred  $i$ -cowit-s look like.

We present four lemmas that allow us to perform this task.

**Lemma 7.** Suppose that  $d \in C_-$  is odd and, for all  $j \leq k$ , either  $b_j = -$  or  $b_j > d$ . If we set  $c_i = b_i$  for all  $i \leq k$ , then  $\mathbf{c}$  is a color witness for  $\rho'$ .

*Proof.* Since  $d < b_j$  for all  $j$ , the outer domination of every  $e$ -color witness implied by  $\mathbf{b}$  is not changed. Moreover, no other

<sup>4</sup>  $k = \lfloor \log_2(e) \rfloor$  again suffices, where  $e$  is the number of vertices with an even color.

property of any  $e$ -color witness is changed by the inclusion of  $v_{m+1}$  in the initial sequence, thus, by setting  $\mathbf{c} = \mathbf{b}$ , we obtain a color witness for  $\rho'$ .  $\square$

Note that the proof of Lemma 7 does not use that  $d$  is odd and holds similarly when  $d$  is even; however, in that case, Lemma 10 provides a better update for the color witness.

**Lemma 8.** Suppose that  $d \in C^-$  is odd, and there exists an index  $j$  such that  $b_j \neq \perp$ ,  $d \geq b_j$ , and, for all  $i > j$ , either  $b_i = \perp$  or  $b_i > d$  hold. Then setting:

- $c_i = b_i$  for all  $i > j$ ,
- $c_j = d$  if  $j \neq 0$  and  $c_j = \perp$  if  $j = 0$ , and
- $c_i = \perp$  for all  $i < j$

yields a color witness for  $\rho'$ .

*Proof.* For all  $e > d$ , the  $e$ -color witness (if any) implied by  $\mathbf{b}$  can be kept: the outer domination of every such  $e$ -color witness implied by  $\mathbf{b}$  is not changed. Moreover, no other property of any such  $e$ -color witness is changed by the inclusion of  $v_{m+1}$  in the initial sequence.

For the  $b_j$ -color witness, we either update the last vertex to  $m + 1$  (if  $b_j$  is odd) or append  $m + 1$  to it (if  $b_j$  is even). In both cases, the inner domination rules are valid (due to the inner and outer domination rules for the  $b_j$ -color witness) and the outer domination rule holds trivially. Moreover, the side constraints for the length carry over from those for  $\mathbf{b}$  (when  $b_j$  is odd), for  $\ell_d$ , by adding one to the length constraint while also appending one state (when  $b_j$  is even).

Thus,  $\mathbf{c}$  is a color witness for  $\rho'$ .  $\square$

**Lemma 9.** Suppose that  $d$  is even, there exists a maximal index  $j$  such that  $b_j < d$ , and  $b_j$  is odd. Then setting:

- for all  $i \geq j$ ,  $c_i = d$  if  $b_i < d$  and  $c_i = b_i$ , otherwise,
- for all  $j > i \geq 1$ ,  $c_i = \perp$ , and
- $c_0 = d$ .

yields a color witness for  $\rho'$ .

*Proof.* We simply append all  $i$ -cowit-s that exist for  $\mathbf{b}$  in the interval  $i \in \{b_j, \dots, d\}$ . We append them in the given order (from the largest  $i$  to the lowest,  $b_j$ ), and then replace the last index (which is from the  $b_j$ -covit) by  $m + 1$ .

The inner domination rules are valid (due to the inner and outer domination rules for the  $i$ -cowit-s involved, and by  $b_j < d$ ). The outer domination rule trivially holds.

The only new rule to be considered is the rule on the joint length of the  $i$ -cowit-s in  $\text{ubp}(d, \mathbf{c})$ , but this is the same length (as only the last element is changed) and the same constraint as for the sum of the length of the  $i$ -cowit-s in  $\text{ubp}(b_j, \mathbf{b})$ .  $\square$

**Lemma 10.** Suppose that  $d$  is even and there is no index  $j'$  such that  $b_{j'} < d$  and  $b_{j'}$  are odd. Let  $j$  be the maximal index (which might be 0) such that:

- for all  $i > j$ ,  $b_i$  is even,  $b_i = \perp$  or  $b_i > d$ ;
- either  $b_j = \perp$  or  $b_j > d$  and  $b_j$  is odd; and
- for all  $i < j$ ,  $b_i$  is even.

If we set:

- $c_i = b_i$  for all  $i > j$  with  $b_i > d$  or  $b_i = \perp$
- $c_i = d$  for all  $i > j$  with  $b_i \leq d$  (and thus even),
- $c_j = d$ , and
- for all  $i < j$ ,  $b_i = \perp$

then  $\mathbf{c}$  is a color witness for  $\rho'$ .

*Proof.* We simply append all  $i$ -cowit-s that exist for  $\mathbf{b}$  in the interval  $i \in \{2, \dots, d\}$ . We append them in the given order (from the largest  $i$  to the lowest), and then append  $m + 1$ .

The inner domination rules are valid (due to the inner and outer domination rules for the  $i$ -cowit-s involved. The outer domination rule trivially holds.

The only new rule to be considered is the rule on the joint length of the  $i$ -cowit-s in  $\text{ubp}(d, \mathbf{c})$ , but this is one more than the length (as only the last element is appended) and the same constraint as for the sum of the length of the  $i$ -cowit-s in  $\text{ubp}(2, \mathbf{b})$ .  $\square$

Again, if  $d = \max\{C\}$  and odd, then the raw update of the color witness is  $\perp \dots \perp$ , which is a color witness for every play prefix.

When we want to update a witness upon scanning another state  $v_{m+1}$  with color  $d = \phi(v_{m+1})$ , we can apply the update rule from one of Lemmas 7 through 10.

For a given witness  $\mathbf{b}$  and a vertex  $v_{m+1}$ , we denote with

- **Raw update:**  $\text{ru}_+(\mathbf{b}, d)$  the raw update of the witness to  $\mathbf{c}$ , as obtained by the update rules described above.
- **Update:**  $\text{up}_+(\mathbf{b}, d)$  is either  $\text{ru}_+(\mathbf{b}, d)$  if  $\text{value}(\text{ru}(\mathbf{b}, d)) \leq e$  (where  $e$  is the number of vertices with even color), or  $\text{up}_+(\mathbf{b}, v_{m+1}) = \text{won}$  otherwise.

In particular,  $\text{up}_+(\text{won}, d) = \text{won}$  for all  $d \in C$ .

- **Antagonistic update:**  $\text{au}_+(\mathbf{b}, v) = \min_{\sqsubseteq} \{\text{up}_+(\mathbf{c}, v) \mid \mathbf{b} \sqsubseteq \mathbf{c} \in \mathbb{C}\}$ .

We first observe that  $\text{up}_+$  is indeed “faster” than  $\text{up}$  in that it always leads to a better (with respect to  $\sqsubseteq$ ) state:

**Lemma 11.** For all  $\mathbf{b} \in \mathbb{C}$  and all  $d \in C$ ,  $\text{up}_+(\mathbf{b}, d) \sqsupseteq \text{up}(\mathbf{b}, d)$ .  $\square$

This is easy to check by the raw update rules, and it entails:

**Corollary 4.** For all  $\mathbf{b} \in \mathbb{C}$  and all  $d \in C$ ,  $\text{au}_+(\mathbf{b}, d) \sqsupseteq \text{au}(\mathbf{b}, d)$ .  $\square$

**Theorem 4.** The following three claims are equivalent for both, forward and backward witnesses:

1. player even has a strategy to win the parity game,
2. player even has a strategy to win the fast basic update game (using  $\text{up}_+$ ), and
3. player even has a strategy to win the fast antagonistic update game (using  $\text{au}_+$ ).

*Proof.* (1) implies (3): By Theorem 3, that player even wins the parity game entails that he wins the concise antagonistic update game. As  $\text{au}_+$  provides (not necessarily strictly) better updates (with respect to  $\sqsubseteq$ ) than  $\text{au}$ , and by the antagonistic update being monotone by definition, this entails (3).

(3) implies (2): as  $\text{up}_+$  provides (not necessarily strictly) better updates (with respect to  $\sqsubseteq$ ) than  $\text{au}_+$  and  $\text{au}_+$  is monotone when  $\text{au}^+$  produces a winning sequence, so does  $\text{up}^+$ .

(2) implies (1): when  $\text{up}^+$  produces a win if, and only if,  $\text{ru}^+$  produces a color witness with value  $> e$ , which according to Lemmas 7 through 10 entails that it has an even chain whose length is strictly greater than  $e$ . The play  $\rho$  must, at that point, contain a cycle, since there must be a vertex with even color that has been visited twice. Moreover, the largest priority on this cycle must be even, so this is a winning cycle for the player even.  $\square$

**Corollary 5.** For a parity game with  $e$  states of even color, colors  $C$ ,  $k = \lfloor \log_2(e) \rfloor$  and  $\mathbb{C}$  the space for concise witnesses of value  $\leq e$ , length  $k+1$  and colors  $C$ , both  $\mathcal{U} = (\mathbb{C}; C; \sqsubseteq, \dots, \sqsubseteq; \text{up}_+; \text{won})$  and  $\mathcal{A} = (\mathbb{C}; C; \sqsubseteq, \dots, \sqsubseteq; \text{au}_+; \text{won})$  are separating automata.  $\square$

To give intuition to their states, for  $\mathcal{U}$  being in a state  $\mathbf{b} \neq \text{won}$  means that  $\mathbf{b}$  is a color witness for the play prefix, while  $\text{won}$  means that the play prefix contains an even chain of length  $> e$ , and thus an even cycle.

For  $\mathcal{A}$  being in a state  $\mathbf{b}$  means that there is a state  $\mathbf{c} \sqsupseteq \mathbf{b}$  with this property.

## 5.4. Faster conversion

While the method will speed up the progress to the solution made by the update operations when using  $\mathcal{A} = (\mathbb{C}; C; \sqsubseteq, \dots, \sqsubseteq; \text{au}_+; \text{won})$  a little, the difference is easier to see when using  $\mathcal{U} = (\mathbb{C}; C; \sqsubseteq, \dots, \sqsubseteq; \text{up}_+; \text{won})$ .

The classic QP algorithms (Calude et al., 2017; Jurdziński and Lazic, 2017; Fearnley et al., 2019) have very simple pathological examples. For example, Jurdziński and Lazic (2017) would traverse the complete statespace for a state with color 2 and a selfloop (or for a state with color 1 and a selfloop, depending on whose player's side the algorithm takes). Similarly, Calude et al. (2017) would traverse very large parts of its statespace when fed with only even colors.

Using our update rules for color witnesses, a loop with even colors will always lead to acceptance within  $e + 1$  steps.

It is possible to make the update rules a bit more robust against the occurrence of odd priorities that are then immediately followed by higher even priorities by returning to  $\mathbb{W}$  as a statespace<sup>5</sup>.

## 6. Statespace

In this section, we compare the size of the statespace with both the statespace from the construction of Jurdziński and Lazic (Jurdziński and Lazic, 2017)—which comes with the best current bounds— and the original statespace from Calude et al. (2017).

We then discuss the effect of the five improvements over the original approach from Calude et al. (2017):

1. the restriction of the number of occurrences of odd colors in a witness to once,
2. not using any color that is higher than any even color;
3. not allowing for odd colors in the rightmost position (i.e.,  $b_0$ );
4. the removal of the color 1; and
5. moving from length to value restriction.

The first of these improvements are, individually, the most powerful one. Three of the other improvements, (2), (3), and (4), have already been discussed in this form in Fearnley et al. (2019).

We will show in Subsection 6.3.1 that applying *only* improvements (1)—the progression from  $\mathbb{W}$  to  $\mathbb{C}$  from Section 4—and (2) leads to a statespace of exactly the same size as that of Jurdziński and Lazic (2017).

Consequently, further improvements, (3)–(5), lead to a strictly smaller statespace. The improvement from (3) alone almost halves the statespace, while (4) alone has only a small effect. The effect of rule (5) — which refers to the change of the functions `evenodd` and `value` described in Section 3.1.2.3 — varies greatly: it is strongest when the bound on the length of an even chain is a power of 2 ( $2^p$  for some  $p \in \mathbb{N}$ ), where it leads to halving the statespace, and vanishes if it is one less ( $2^p - 1$  for some  $p \in \mathbb{N}$ ).

After briefly visiting the statespace from Fearnley et al. (2019), we then turn to an experimental comparison of the three statespaces of interest, confirming the quantification of the advantage we have obtained over (Jurdziński and Lazic, 2017).

<sup>5</sup> When using  $\mathbb{W}$ , there would need to be some care taken with odd  $\sigma$ -witnesses: while the rules for overwriting lower numbers are as expected, the point to bear in mind that the treatment of odd colors that already occur in a witness (covered by Lemma 8) generalize to if there is already a lowest position  $j$  with  $b_j = \sigma$ , then just replace all  $b_i$  with  $i < j$  by  $\perp$ . This is because, while two even color witnesses can be merged, two odd color witnesses cannot, and there would be no means to mark them as different color witnesses.

In this section, we use  $\text{count}_{alg}^{size}$  for counting the number of state minus one, estimating the number of states except for the winning state (“won”), which all progress measures under consideration have. The superscript *size* can be  $\ell$ , saying that only the length of the data structure (or: the  $\lceil \log_2(e+1) \rceil$  for the maximal length  $e$  of an even chain) is taken into account;  $v$  if the value of witness is taken into account (or: the maximal length  $e$  of an even chain) is taken into account, and  $\ell, v$  if both are used. The subscript is either  $JL$  when counting the concise progress measures from Jurdziński and Lazic (2017),  $O$  when considering the original approach from Calude et al. (2017), ‘1, 2’ when adding improvements (1) and (2), or blank when considering either improvement (1) through (4) or all improvements. In a closing comparison with the statespace of Fearnley et al. (2019), we use the subscript  $JKSSW$ .

## 6.1. Concise progress measures

We will not describe the algorithm, but the data structure, which holds a winning state besides the states we describe. For each even priority, there is a (possibly empty) word over two symbols, say  $+$  and  $-$ , such that the words concatenated have length at most  $\ell = \lceil \log_2(e+1) \rceil$ , where  $e$  is the number of states with an even priority. That is,  $\ell$  is the length of the witness and color witness from the previous section ( $\ell = k+1$ ).

For Jurdziński and Lazic (2017) (i.e.,  $alg = JL$ ), with  $c$  priorities  $\{1, \dots, c\}$  and  $n$  states with even priority (not counting the winning state) we have the following counts:

- Induction basis, length: we start with the case in which we bound the sum of the lengths of  $+$  and  $-$  by 0 or 1.

When we bound the sum of the lengths by 0, then there is only one sequence:

$$\text{count}_{JL}^{\ell}(c, 0) = 1,$$

and when we bound it by 1, then we get:

$$\text{count}_{JL}^{\ell}(2c, 1) = \text{count}_{JL}^{\ell}(2c+1, 1) = 2c+1,$$

as there are  $c$  positions in which a sequence of length 1 can occur (one for each even priority in  $\{1, \dots, 2c+1\}$  or  $\{1, \dots, 2c\}$ , respectively), and there is one for the case in which all sequences have length 0.

- Induction basis, colors: when there is only one even color (i.e., 2), we have:

$$\text{count}_{JL}^{\ell}(3, l) = \text{count}_{JL}^{\ell}(2, l) = 2^{l+1} - 1.$$

These are the binary words of length at most  $l$ .

- For all other cases, we define inductively:

$$\begin{aligned} \text{count}_{JL}^{\ell}(2c+1, l) &= \text{count}_{JL}^{\ell}(2c, l) = \text{count}_{JL}^{\ell}(2c-2, l) \\ &+ 2\text{count}_{JL}^{\ell}(2c, l-1), \end{aligned}$$

where the first summand refers to the case where the leading sequence (which refers to color  $2c$ ) is empty. In this case, the length of the remaining sequences is still bound by  $l$ , but the number of even colors has dropped by one. The two summands  $\text{count}_{JL}^{\ell}(2c, l-1)$  represent the cases, where the sequence assigned to the highest even priority starts with a  $+$  and  $-$ , respectively. Cutting off this leading sign leaves  $\text{count}_{JL}^{\ell}(2c, l-1)$  in different states.

To estimate the number of concise progress measures,  $\text{count}_{JL}^{\ell}(c, l) + 1$ , we put aside the winning state and the function that maps all even priorities to the empty sequence.

For the remaining states, we first fix a positive length  $i \leq l$  of the concatenated words, and then the  $j \leq \lfloor c/2 \rfloor$  of the even priorities that have a non-empty word assigned to them.

There are  $\binom{i-1}{j-1}$  assignments of positive lengths to  $j$  positions that add up to  $i$ . For each distribution of lengths, there are  $2^i$  different assignments to words. Finally, there are  $\binom{\lfloor c/2 \rfloor}{j}$  possibilities to assign  $j$  of the  $\lfloor c/2 \rfloor$  different even priorities.

This provides an overall statespace of

$$2 + \sum_{i=1}^l \sum_{j=1}^{\min\{i, \lfloor c/2 \rfloor\}} 2^i \cdot \binom{\lfloor c/2 \rfloor}{j} \cdot \binom{i-1}{j-1}.$$

## 6.2. Calude et al.

‘We now continue with the statespace of the original quasi-polynomial approach of Calude et al. (2017), hence,  $alg = O$ . The statespace used in Calude et al. (2017) is slightly larger than  $\mathbb{W}$ , as it only uses the length of the witness as a restriction and does not exclude odd values for the rightmost position (“ $b_0$ ”) in a witness.

For the precise count of this statespace without the winning state “won,” we have the following counts:

- Induction basis, length: sequences of length 1 (only containing  $b_0$ ) whose color values are bounded by  $c$ , can take  $c+1$  values in  $\{1, \dots, c\}$  plus  $-$ . We, therefore, have:

$$\text{count}_O^{\ell}(c, 1) = c+1.$$

- For longer tails of sequences, we define inductively:

$$\text{count}_O(c, l+1) = \text{count}_O(c, l) + \sum_{i=1}^c \text{count}_O(i, l).$$



The summands refers to the possible values taken by the leftmost position (“ $b_l$ ”) of the tail  $(b_l, b_{l-1}, \dots, b_0)$ . The first summand refers to the leading position being “ $b_l = \_$ ”. This does not restrict the values the rest of the tail may take any further as the highest color allowed to appear is still  $c$ . The other summands refer to the value  $i$  (“ $b_l = i$ ”). When the value of the leftmost position is  $i \leq c$ , then the highest color that may occur in the remaining positions is  $i$ .

The size of  $|\mathbb{W}^+|$  of the statespace can be given as:

$$2 + \sum_{i=1}^l \binom{l}{i} \cdot \binom{i+c-1}{i}.$$

The “2” refers to the winning state and the “empty” sequence consists only of  $\_$  symbols (which is more convenient for us to treat separately), and the sum refers to the states represented by non-empty sequences of length  $l = \lceil \log_2(e+1) \rceil$ , where  $e$  is the number of states with an even priority. Note that the estimation given in Calude et al. (2017) is slightly coarser, and their game definition is slightly different from the normal definition of parity games, but the bound can be taken from Fearnley et al. (2019).

For the estimation, after fixing the positive  $i \leq l$  positions with values different to  $\_$ , there are  $\binom{i+c-1}{i}$  different valuations when we have  $c$  priorities. For each  $i \leq l$ , there are  $\binom{l}{i}$  different choices for the  $i$  positions containing some number  $d \in C$ . This leads to  $\sum_{i=1}^l \binom{l}{i} \cdot \binom{i+c-1}{i}$  different states that contain  $i \leq l$  positions that have a value in  $C$ .

To obtain a better inroad to outline the differences, we first take a closer look at the  $\binom{i+c-1}{i}$  different valuations that we may have for  $c$  priorities when we have fixed the  $i \leq l$  positions that are not marked as  $\_$ .

For these positions, we can look at the cases where there are  $j \leq \min(i, r)$  fixed different priorities. For that case, there are  $\binom{i-1}{j-1}$  many assignments of these  $j$  priorities to the  $i$  positions. Moreover, there are  $\binom{c}{j}$  different options to select  $j$  of the available  $r$  priorities. Thus, we get

$$\binom{i+c-1}{i} = \sum_{j=1}^{\min\{i,c\}} \binom{c}{j} \cdot \binom{i-1}{j-1}$$

different combinations for all number of priorities put together, providing the following size:

$$2 + \sum_{i=1}^l \sum_{j=1}^{\min\{i,c\}} \binom{l}{i} \cdot \binom{c}{j} \cdot \binom{i-1}{j-1}.$$

## 6.3. Improvements

We now discuss the differences obtained when moving from  $\mathbb{W}^+$  to  $\mathbb{C}$  by looking at the effect of the three optimisations we have introduced. These are:

1. the restriction of the number of occurrences of odd colors in a witness to once,
2. not using any color that is higher than any even color;
3. not allowing for odd colors in the rightmost position (“ $b_0$ ”);
4. the removal of the color 1; and
5. moving from length to value restriction.

### 6.3.1. (1) and (2) Restricted occurrence of odd colors

Restricting the occurrence of odd colors to once, together with the optimization of not using any color that is higher than any even color, leads to a situation where the highest color allowed in any position is even. To see this, we observe that the banning of a potential odd color higher than any even color guarantees this initially, where the highest color allowed is the highest even color.

When an odd color  $o$  is used in the witness (“ $b_l = o$ ”), then the highest color allowed to its right is  $o-1$ , whereas when an even color  $e$  is used in the witness (“ $b_l = e$ ”), then the highest color allowed to its right is  $e$ .

We, therefore, only have to define our improved counting function for even colors:

- Induction basis, length: apart from using only even bounds, the base case remains the same:

$$\text{count}_{1,2}^{\ell}(2c, 1) = 2c + 1.$$

- For longer tails of sequences, we define inductively:

$$\text{count}_{1,2}^{\ell}(2c, l+1) = 1 + 2 \sum_{i=1}^c \text{count}_{1,2}^{\ell}(2i, l).$$

The summands refer to the possible values taken by the leftmost position (“ $b_l$ ”) of the tail  $(b_l, b_{l-1}, \dots, b_0)$ .

The first summand refers to the leading position being 1 (“ $b_l = 1$ ”). If this is the case, then all entries to its right must be *strictly smaller* than 1 (which is not possible) or  $\_$ —consequently, they must all be  $\_$ , which just leaves one such tail.

The other summands refer to the leading position taking the value  $2i$  or  $2i + 1$  when  $i < c$  (" $b_l = 2i$ " or " $b_l = 2i+1$ "), in either case, the maximal value of the colors occurring in the remaining tale is  $2i$ .

The final two summands (for  $i = c$ ) refer to the leading position taking the value  $2c$  or  $\_$  (" $b_l = 2c$ " or " $b_l = \_$ "). In both cases, the maximal value of the colors occurring in the remaining tale is  $2c$ .

While the representation is different, it is easy to see that  $\text{count}_{1,2}^\ell(2c, l) = \text{count}_{JL}^\ell(2c, l)$  holds.

To see this, we first observe that  $\text{count}_{JL}^\ell(2, l+1) = 1 + 2\text{count}_{JL}^\ell(2, l)$  holds, and then by induction over  $c$  that:

$$\text{count}_{JL}^\ell(2, l+1) = 1 + 2 \sum_{i=1}^c \text{count}_{JL}^\ell(2i, l).$$

Given that we also have  $\text{count}_{1,2}^\ell(2c, 1) = \text{count}_{JL}^\ell(2c, 1)$ , we get the claim, because  $\text{count}_{JL}^\ell(2c, l)$  cannot be derived in the same way as  $\text{count}_{1,2}^\ell(2c, l)$ .

### 6.3.2. (3) and (4) Removing odd colors from the rightmost position and 1s

Removing odd colors from the rightmost positions only changes the base case, while banning 1 from the other positions merely removes the "1+" part from the inductive definition. This leaves:

- Induction basis, length:

$$\text{count}^\ell(2c, 1) = c + 1.$$

- For longer tails of sequences, we define inductively:

$$\text{count}^\ell(2c, l+1) = 2 \sum_{i=1}^c \text{count}^\ell(2i, l).$$

When evaluating the term  $\text{count}^\ell$ , the reduction from  $2c+1$  to  $c+1$  is halving the value (rounded up) at the leaf of each call tree, which provides more than the removal of " $= 1$ " in each node of the call tree. Together, they **broadly halve the value**.

### 6.3.3. (5) Taking the value into account

We start with using both the length and the value and then remove the length in the next step to get a more concise representation, but we note that, for a given length  $l$ , the value  $v$  allowed always satisfies  $v < 2^l$ .

First, we get another induction base, one by value:

- Induction basis, value:

$$\text{count}^{\ell,v}(2c, l, 0) = 1.$$

Regardless of the remaining length, if the *value* of the tail is bounded by (and thus needs to be) 0, then it can only consist of  $\_$  signs.

- Induction basis, length:

$$\text{count}^{\ell,v}(2c, 1, 1) = c + 1.$$

- For longer tails of sequences and positive values, we distinguish several cases. The first case is that  $v < 2^l$ . Then we have

$$\text{count}^{\ell,v}(2c, l+1, v) = \text{count}^{\ell,v}(2c, l, v).$$

This is simply because filling the position  $l+1$  with any number, even or odd, would exceed the value budget.

This leaves the case  $v \geq 2^l$ , i.e.,:

$$\begin{aligned} \text{count}^{\ell,v}(2c, l+1, v) = & \sum_{i=1}^c \text{count}^{\ell,v}(2i, l, v - 2^l) \\ & + \sum_{i=1}^c \text{count}^{\ell,v}(2i, l, 2^l - 1). \end{aligned}$$

This is because, when filling position  $l$  with an even number, it takes  $2^l$  from the budget of the value, leaving a remaining budget of  $v - 2^l$ .

When filling this position with an odd number, while the value would be increased by  $2^l$ , this is within the value budget. Moreover, if this position is still relevant to the value, then the positions to its right no longer add to the value of the sequence, as the leftmost odd position would be the last to be considered.

We, therefore, set the value for the remaining tail to the right to be the maximal value that can be obtained by this tail, which is  $2^l - 1$ ; this is a rendering of saying that for the tail the values are not constrained.

The effect of adding value can vary greatly. It is larger when the number of positions with even color is a power of 2, say  $2^l$ , and it has no effect at all if it is  $2^2 - 1$ . In the former case, if the initial position is even, then all other positions need to be  $\_$ . Generally, we have

$$\begin{aligned} \text{count}^{\ell,v}(2c, l, 2^l - 1) &= \text{count}^\ell(2c, l) \quad \text{and} \\ \text{count}^{\ell,v}(2c, l, 2^{l-1}) &= \text{count}^\ell(2c, l)/2 + c \end{aligned}$$

for all  $l > 1$ .

Taking the value into account therefore broadly **halves the statespace when  $e$  is a power of 2**, and **has no effect when  $e$  is**

a predecessor of a power of 2, falls from  $2^{l-1}$  to  $2^l - 1$  for all  $l > 1$ .

Looking at the definition of  $\text{count}^{\ell,v}$ , it is easy to see that an explicit reference to the length can be replaced by a reference to the next relevant length,  $\lfloor \log_2 v \rfloor$ . This provides:

$$\begin{aligned}\text{count}^v(2c, 0) &= 1, \\ \text{count}^v(2c, 1) &= c + 1, \text{ and} \\ \text{count}^v(2c, v) &= \sum_{i=1}^c \text{count}^v(2i, v - 2^{\lfloor \log_2 v \rfloor}) \\ &\quad + \sum_{i=1}^c \text{count}^v(2i, 2^{\lfloor \log_2 v \rfloor} - 1) \text{ otherwise.}\end{aligned}$$

## 6.4. Comparison with the statespace of

While improvement (1) is the most powerful of the optimizations, the improvements (2)–(4) were present in Fearnley et al. (2019), where the algorithm makes use of a value function, namely  $\text{value}'(\mathbf{b}) = \sum_{i \in \text{even}(\mathbf{b})} 2^i$ . It is, therefore, interesting to provide a count function for Fearnley et al. (2019). We use the subscript JKSSW, and only use the count that uses both length and value.

We get the following state counts:

$$\begin{aligned}\text{count}_{JKSSW}^{\ell,v}(c, 1, 0) &= 1 \\ \text{count}_{JKSSW}^{\ell,v}(c, 1, 1) &= \lfloor c/2 \rfloor + 1 \\ \text{if } v < 2^l: \text{count}_{JKSSW}^{\ell,v}(c, l+1, v) &= \text{count}_{JKSSW}^{\ell,v}(c, l, v) \\ &\quad + \sum_{i=2}^{\lceil c/2 \rceil} \text{count}_{JKSSW}^{\ell,v}(2i-1, l, v) \\ \text{if } v \geq 2^l: \text{count}_{JKSSW}^{\ell,v}(c, l+1, v) &= \text{count}_{JKSSW}^{\ell,v}(c, l, 2^l-1) \\ &\quad + \sum_{i=1}^{\lfloor c/2 \rfloor} \text{count}_{JKSSW}^{\ell,v}(2i, l, v-2^l) \\ &\quad + \sum_{i=2}^{\lceil c/2 \rceil} \text{count}_{JKSSW}^{\ell,v}(2i-1, l, 2^l-1).\end{aligned}$$

To explain the difference to  $\text{count}^{\ell,v}$ , one major difference is that the highest color allowed in a position can be odd. The other is that positions with odd color do not contribute to the weight, which allows for adding positions with odd color the remaining budget is lower than  $2^l$ .

Thus, the call tree for the calculation of  $\text{count}_{JKSSW}^{\ell,v}$  has  $\lfloor c/2 \rfloor$  successors where  $v < 2^l$  while the call tree for  $\text{count}^{\ell,v}$  has just one. For  $v \geq 2^l$ , the call tree has the same number of successors (for even  $c$ ) or just one additional successor (for odd  $c$ ), but the parameter falls slower.

## 7. Statespace comparison

In this subsection, we provide a graphical representation of the statespace size for the three algorithms: Calude et al. (2017), Jurdziński and Lazic (2017), and the improvement described in this article. The size of the statespace on which

an algorithm works does not represent how well the algorithm performs in practice. Indeed, in the context of parity games, there are quasi-polynomial time algorithms that behave like brute-force approaches. Therefore, they always require quasi-polynomial many steps to compute the solution, while most of the exponential time algorithms, instead, almost visit a polynomial fraction of their statespace. The first improvement we described does not affect the performance of the algorithm, since both the original and the improved algorithm require the same number of steps to solve a game, but the latter works on a reduced statespace. To measure how big is the cut we consider Figure 1 games with a fixed number of colors and Figure 2 games with a linear number of colors in the size of the game. The games in Figure 1 range from  $2^3$  to  $2^{15}$  positions  $n$ . Therefore, the length of the measure, which is logarithmic in  $n$ , constantly increases, while the colors are fixed to a value of 10. As a consequence, the ratio of colors with respect to  $n$  range from 80 to 0.02%. As expected, the cut with the original algorithm significantly increases for games that are not dense in colors as the lines tend to diverge on a logarithmic scale. The ratio between Jurdzinski and Lazic approach (JL) and the new improvement, instead, converges to a cut of 73% of the statespace. The games of Figure 2, instead range from  $2^8$  to  $2^9$  positions  $n$ , so that the length of the measure is fixed, while the number of colors constantly grows from 26 to 50. As a consequence, we have that the ratio of colors with respect to  $n$  is fixed to 10%. In this case, the scale is linear and, even if the improved statespace is always smaller than the other two, the cut tends to shrink.

## 8. Discussion

We have introduced three technical improvements over the progress measures used in the original quasipolynomial approach by Calude et al. (2017) and its improvements by Fearnley et al. (2019). The first two reduce the statespace.

The more powerful of the two is a simple limitation of the occurrences of odd colors in a witness to one. Where the highest color is even, this alone reduces the size of the statespace of Calude et al.'s approach to the currently smallest one of Jurdziński and Lazic (2017). Where the highest color is odd, we obtain the same by borrowing the simple observation that this highest color does not need to be used from Fearnley et al. (2019).

The second new means to reduce the statespace is the use of witnesses that only refer to even chains of plausible size, namely those that do not contain more dominating even states than the game has to offer. A similar idea had been explored in Fearnley et al. (2019), but our construction is more powerful in reducing the size of the statespace. The effect of this step ranges from none (where the number of states with even color is the predecessor of a power of 2 ( $2^\ell - 1$  for some  $\ell \in \mathbb{N}$ ), then rises steeply to a

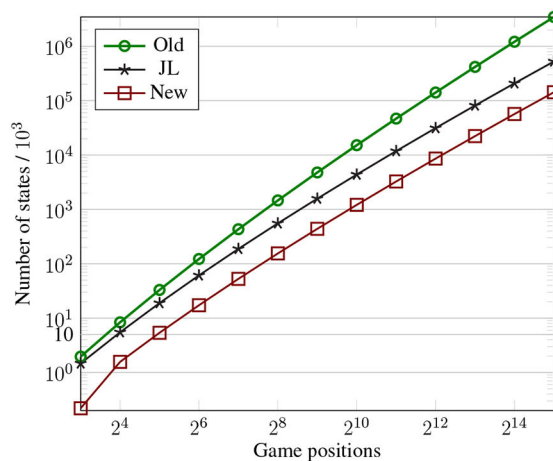


FIGURE 1  
Size of the statespace for games with a fixed number of colors on a logarithmic scale.

Game data		Statespace size / $10^3$		
Nodes	Colours	Old	JL	New
8	8	2	1	>1
16	10	8	5	1
32	10	33	18	5
64	10	122	61	17
128	10	432	187	52
256	10	1462	553	154
512	10	4780	1579	439
1024	10	15157	4374	1211
2048	10	46813	11829	3261
4096	10	141264	31326	8601
8192	10	417577	81461	22282
16384	10	1211700	208470	56819
32768	10	3458200	525991	142884

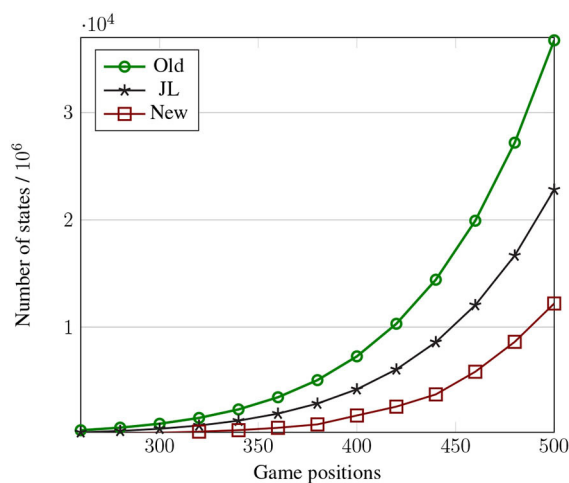


FIGURE 2  
Size of the statespace for games with a linear number of colors on a linear scale.

Game data		Statespace size / $10^6$		
Nodes	Colours	Old	JL	New
260	26	381	190	53
280	28	622	318	90
300	30	987	518	148
320	32	11531	820	251
340	34	2323	1271	389
360	36	3456	1928	608
380	38	5054	2870	926
400	40	7271	4201	1759
420	42	10309	6053	2584
440	44	14420	8596	3724
460	46	19919	12047	5838
480	48	27199	16675	8625
500	50	36742	22818	12200

factor of 2 for a power of 2 ( $2^\ell$ ), and then slowly falls again, until it vanishes at the next predecessor of a power of 2.

These improvements work well with the other improvements from Fearnley et al. (2019), namely not using the color 1 and disallowing odd values for the rightmost position (“ $b_0$ ”) in a witness. These improvements broadly halve the statespace, leading to a statespace reduction that broadly oscillates between 2 and 4 when compared to the previously leading approach.

The second improvement we have introduced is a re-definition of the semantics of witnesses, moving from classic *witnesses* to *color witnesses*. While it does not lead to a difference in the size of the statespace, it does accelerate its traversal,

especially for the “standard” update rule that does not extend to value iteration; in particular, it gets rid of the most trivial kind of silly hard examples, such as cliques of states of player odd that all have even color.

While the use of *color witnesses* clearly accelerates the analysis, it is not as easy as for the statespace reduction to formally quantify this advantage.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary




material, further inquiries can be directed to the corresponding author.

## Author contributions

Both authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101032464 .

## References

- Alur, R., Henzinger, T., and Kupferman, O. (2002). Alternating-time temporal logic. *J. ACM* 49, 672–713. doi: 10.1145/585265.585270
- Benerecetti, M., Dell'Erba, D., and Mogavero, F. (2016). "Improving priority promotion for parity games," in *HVC'16, LNCS 10028* (Haifa: Springer), 1–17.
- Benerecetti, M., Dell'Erba, D., and Mogavero, F. (2018a). A delayed promotion policy for parity games. *Inf. Comput.* 262, 221–240. doi: 10.1016/j.ic.2018.09.005
- Benerecetti, M., Dell'Erba, D., and Mogavero, F. (2018b). Solving parity games via priority promotion. *Form. Methods Syst. Des.* 52, 193–226. doi: 10.1007/s10703-018-0315-1
- Benerecetti, M., Mogavero, F., and Murano, A. (2013). "Substructure temporal logic," in *Logic in Computer Science'13* (New Orleans, LA: IEEECS), 368–377.
- Berwanger, D., and Grädel, E. (2004). Fixed-point logics and solitaire games. *Theor. Comput. Sci.* 37, 675–694. doi: 10.1007/s00224-004-1147-5
- Bojańczyk, M., and Czerwinski, W. (2018). Available online at: <https://www.mimuw.edu.pl/~bojan/papers/toolbox-reduced-feb6.pdf> An Automata Toolbox.
- Calude, C., Jain, S., Khossainov, B., Li, W., and Stephan, F. (2017). "Deciding parity games in quasipolynomial time," in *Symposium on Theory of Computing'17* (Montreal, QC: Association for Computing Machinery), 252–263.
- Chatterjee, K., Henzinger, T., and Piterman, N. (2010). Strategy logic. *Inf. Comput.* 208, 677–693. doi: 10.1016/j.ic.2009.07.004
- Czerwinski, W., Daviaud, L., Fijalkow, N., Jurdzinski, M., Lazic, R., and Parys, P. (2018). "Universal trees grow inside separating automata: quasi-polynomial lower bounds for parity games," in *SODA'18* (SIAM), 2333–2349.
- Emerson, E., and Jutla, C. (1991). "Tree automata, mu-calculus, and determinacy," in *FOCS'91* (San Juan: IEEECS), 368–377.
- Emerson, E., Jutla, C., and Sistla, A. (2001). On model checking for the mu-calculus and its fragments. *Theor. Comput. Sci.* 258, 491–522. doi: 10.1016/S0304-3975(00)00034-7
- Emerson, E., and Lei, C.-L. (1986). "Temporal reasoning under generalized fairness constraints," in *Symposium on Theoretical Aspects of Computer Science'86, LNCS 210* (Orsay: Springer), 267–278.
- Fearnley, J. (2010). "Non-oblivious strategy improvement," in *LPAR'10, LNCS 6355* (Dakar: Springer), 212–230.
- Fearnley, J., Jain, S., Keijzer, B., Schewe, S., Stephan, F., and Wojtczak, D. (2019). An ordered approach to solving parity games in quasi polynomial time and quasi linear space. *Software Tools Technol. Transfer* 21, 325–349. doi: 10.1007/s10009-019-00509-3
- Fearnley, J., Jain, S., Schewe, S., Stephan, F., and Wojtczak, D. (2017). "An ordered approach to solving parity games in quasi polynomial time and quasi linear space," in *SPIN'17* (Santa Barbara, CA: Association for Computing Machinery), 112–121.
- Friedmann, O. (2013). A superpolynomial lower bound for strategy iteration based on snare memorization. *Discrete Appl. Math.* 161, 1317–1337. doi: 10.1016/j.dam.2013.02.007
- Grädel, E., Thomas, W., and Wilke, T. (2002). "Automata, logics, and infinite games: a guide to current research," in *LNCS 2500* (Dagstuhl: Springer).
- Jurdzinski, M. (1998). Deciding the winner in parity games is in  $UP \cap co-UP$ . *Inf. Process. Lett.* 68, 119–124. doi: 10.1016/S0020-0190(98)00150-1
- Jurdzinski, M. (2000). "Small progress measures for solving parity games," in *Symposium on Theoretical Aspects of Computer Science'00, LNCS 1770* (Lille: Springer), 290–301.
- Jurdzinski, M., and Lazic, R. (2017). "Succinct progress measures for solving parity games," in *Logic in Computer Science'17* (Reykjavik: Association for Computing Machinery), 1–9.
- Kupferman, O., and Vardi, M. (1998). "Weak alternating automata and tree automata emptiness," in *Symposium on Theory of Computing'98* (Dallas, TX: Association for Computing Machinery), 224–233.
- Lapauw, R., Bruynooghe, M., and Denecker, M. (2020). "Improving parity game solvers with justifications," in *VMCAI'20, LNCS 11990* (New Orleans, LA: Springer), 449–470.
- Lehtinen, K. (2018). "A modal mu perspective on solving parity games in quasi-polynomial time," in *Logic in Computer Science'18* (Oxford: Association for Computing Machinery & IEEECS), 639–648.
- Martin, A. (1975). Borel determinacy. *Ann. Math.* 102, 363–371. doi: 10.2307/1971035
- Mogavero, F., Murano, A., Perelli, G., and Vardi, M. (2012). "What makes ATL\* decidable? a decidable fragment of strategy logic," in *Concurrency Theory'12, LNCS 7454* (Newcastle upon Tyne: Springer), 193–208.
- Mogavero, F., Murano, A., and Vardi, M. (2010). "Reasoning about strategies," in *FSTTCS'10, LIPIcs 8* (Chennai: Leibniz-Zentrum fuer Informatik), 133–144.
- Mostowski, A. (1991). *Games with Forbidden Positions*. Technical report, University of Gdańsk, Gdańsk, Poland.
- Parys, P. (2019). "Parity games: Zielonka's algorithm in quasi-polynomial time," in *Proceedings of MFCS, LIPIcs 138* (Leibniz-Zentrum fuer Informatik), 1–10.
- Schewe, S. (2007). "Solving parity games in big steps," in *FSTTCS'07, LNCS 4855* (New Delhi: Springer), 449–460.
- Schewe, S. (2008). "ATL\* satisfiability is 2ExpTime-complete," in *International Colloquium on Automata, Languages, and Programming'08, LNCS 5126* (Reykjavik: Springer), 373–385.
- Schewe, S., and Finkbeiner, B. (2006). "Satisfiability and finite model property for the alternating-time mu-calculus," in *CSL'06, LNCS 6247* (Szeged: Springer), 591–605.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

van Dijk, T. (2018). "Attracting tangles to solve parity games," in *CAV'18, LNCS 10982* (Oxford: Springer), 198–215.

Vöge, J., and Jurdziński, M. (2000). "A discrete strategy improvement algorithm for solving parity games," in *CAV'00, LNCS 1855* (Chicago, IL: Springer), 202–215.

Wilke, T. (2001). Alternating tree automata, parity games, and modal  $\mu$ Calculus. *Bull. Belg. Math. Soc.* 8, 359–391. doi: 10.36045/bbms/1102714178

Zielonka, W. (1998). Infinite games on finitely coloured graphs with applications to automata on infinite trees. *Theor. Comput. Sci.* 200, 135–183. doi: 10.1016/S0304-3975(98)00009-7



# Skeletons, Object Shape, Statistics

Stephen M. Pizer<sup>1\*</sup>, J. S. Marron<sup>2</sup>, James N. Damon<sup>3</sup>, Jared Vicory<sup>4</sup>, Akash Krishna<sup>1</sup>, Zhiyuan Liu<sup>1</sup> and Mohsen Taheri<sup>5</sup>

<sup>1</sup> Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, <sup>2</sup> Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States,

<sup>3</sup> Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, <sup>4</sup> Kitware Inc., Carboro, NC, United States, <sup>5</sup> Department of Mathematics and Physics, University of Stavanger, Stavanger, Norway

Objects and object complexes in 3D, as well as those in 2D, have many possible representations. Among them skeletal representations have special advantages and some limitations. For the special form of skeletal representation called “s-reps,” these advantages include strong suitability for representing slabular object populations and statistical applications on these populations. Accomplishing these statistical applications is best if one recognizes that s-reps live on a curved shape space. Here we will lay out the definition of s-reps, their advantages and limitations, their mathematical properties, methods for fitting s-reps to single- and multi-object boundaries, methods for measuring the statistics of these object and multi-object representations, and examples of such applications involving statistics. While the basic theory, ideas, and programs for the methods are described in this paper and while many applications with evaluations have been produced, there remain many interesting open opportunities for research on comparisons to other shape representations, new areas of application and further methodological developments, many of which are explicitly discussed here.

**Keywords:** shape, skeleton, shape statistics, skeletal model, s-reps

## OPEN ACCESS

### Edited by:

Marcello Pelillo,

Ca' Foscari University of Venice, Italy

### Reviewed by:

Anuj Srivastava,

Florida State University, United States

Christoph von Tycowicz,

Freie Universität Berlin, Germany

### \*Correspondence:

Stephen M. Pizer

pizer@cs.unc.edu

### Specialty section:

This article was submitted to

Computer Vision,

a section of the journal

Frontiers in Computer Science

**Received:** 04 February 2022

**Accepted:** 13 June 2022

**Published:** 18 October 2022

### Citation:

Pizer SM, Marron JS, Damon JN,

Vicory J, Krishna A, Liu Z and

Taheri M (2022) Skeletons, Object

Shape, Statistics.

Front. Comput. Sci. 4:842637.

doi: 10.3389/fcomp.2022.842637

## INTRODUCTION

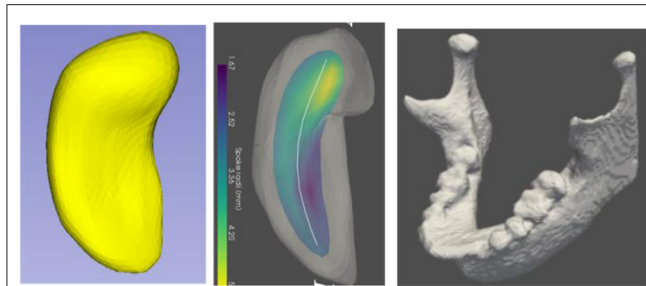
This paper discusses models of objects, in 3D and 2D but with special emphasis on 3D. The concern is with both individual objects and complexes of multiple objects and, especially, *objects and complexes that appear as populations of instances*. The objects of concern have curved slab-like shapes, with no branching or a limited, fixed amount of branching. I<sup>1</sup> call these objects “slabular,” with tube-like objects, i.e., generalized cylinders, being a special case of slabular objects. The examples we will start with are anatomic objects, such as the hippocampus or mandible (jaw-bone) shown in Figure 1<sup>2</sup>, but as will be seen later, the ideas apply to many manufactured objects, such as shoe boxes or airplanes.

The purpose of this paper is to survey a few decades of work on a skeletal model called “s-reps,” designed for statistics. The paper provides many mathematical and statistical concepts, definitions, and properties. It describes many algorithms related to forming and using s-reps, and it describes many uses to which they have been put. It argues that the focus of these models on the objects themselves and capabilities of s-reps in richly capturing geometric properties, especially for statistical applications, make it an important way of describing shape.

The most common alternative approaches to representing shape are point distribution models (PDMs), especially focusing on boundary points, and models describing shape via diffeomorphisms from a containing space enclosing the objects (e.g., Durrleman et al., 2014), as distinct from

<sup>1</sup>In this paper “I” refers to concepts and approaches specifically credited to the first author, while “we” refers to the collection of authors.

<sup>2</sup>Many of the figures in this paper also appear in Pizer et al. (2021).

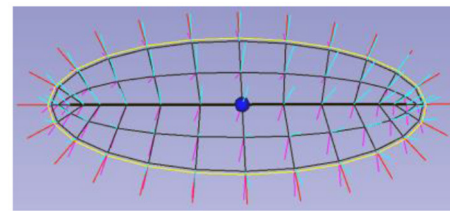


**FIGURE 1** | Left: A hippocampus, a C-shaped object. Middle: the hippocampus with its skeletal surface (colored, with the color indicating the width of the object there) and its spine (the curve that is its long axis). Right: a mandible, another C-shaped object.

s-rep methods described here that are formed by diffeomorphisms of object interiors. The experimental data to date, cited and/or presented in this paper is that s-reps are superior to PDMs for statistical applications and that using s-reps to form PDMs has superiority to other approaches of forming PDMs. We are not aware of experiments comparing s-reps to the methods based on diffeomorphisms of a containing space for statistical purposes, and **we look forward to such comparisons.**

Despite decades of development, in laboratories at or in collaboration with mine at UNC, there are many statistical, mathematical, algorithmic, and applications challenges still to be met, as well as further studies comparing s-reps to other shape representations. **One of the objectives of this paper is to lay out these remaining challenges and opportunities; each of these are indicated in the Bold font.**

For a slabular object there exists a smooth sequence of usually non-parallel slicing planes such that no successive planes intersect within the object and such that the object boundary's intersection with each plane is eccentric (one of the two principal axes' radii is notably longer than the other). As described in Section Skeletal Models and S-reps: Definitions and Mathematics, the locus of centers of the cross-sections forms a curvilinear axis, which I call the "spine," that is notably longer than the axes in the cross-sections. For example, for the hippocampus seen in **Figure 1** the long axis is C-shaped and goes from the tip (at the top of the figure) to the tail, the short axis goes from the front of the object as seen in **Figure 1** left to the back from that point of view, and the middle-length dimension goes from side to side as seen in that Figure. In the mandible the shortest axis goes from the facial side to the inside of the mouth, the longest axis goes from one temporal-mandibular joint (TMJ, where the mandible hinges on the skull) to the other, and the middle-length axis goes from the teeth positions extended to the TMJs down to the chin locus extended to the TMJs. If the longest axis terminates in the two knob-shaped entities (called the "condyles"), the pointy figures opposite the condyles, called the "coronoid processes," form subfigures. These subfigures can be found in essentially every human mandible.



**FIGURE 2** | An ellipsoid's skeleton sampled into a grid and its spokes; spokes at the skeletal fold are displayed as red, those on the north side of the skeleton in cyan, and those on the south side in magenta. The spine is bold, and the center point is displayed as a bullet.

Since all slabular objects have a central curve, formed by the spine, and cross-sections, they can all be understood as generalized cylinders. However, when the cross-sections' two axes are not too different in length, that is, the cross-sections are not too far from circular, the generalized cylinder is more tube-like.

While the most common computer representations of such objects capture either their boundary locations alone or deformations of the whole ambient space in which the objects reside, it has been seen by many (e.g., Blum and Nagel, 1978; Amenta and Choi, 2008; Siddiqi, 2008; Székely, 2008; Yushkevich et al., 2015) that object widths, which are captured by the shorter cross-sectional axes, are important features as are features derivable from the behaviors of specified directions, especially boundary normals (Srivastava et al., 2011). Since these features are exactly what skeletal models capture, in quite a variety of applications such models have been shown more powerful than those based on the other types of object representations. From the point of view of slabular objects the skeleton (**Figure 1**, middle) includes the spine as its long axis, and the cross-sectional dimensions can be captured by line segments emanating at the spine and ending at the edge of the skeleton (**Figure 2**). We call these line segments "spokes".

Given a population of objects skeletally represented, it is useful to model the populations via probability densities. It has been shown that these can be of use for objectives such as classification into subpopulations, hypothesis testing on differences between populations, and segmentation approaches that use shape prior distributions as well as image intensity features understood via the spatial correspondences that the geometric model provides. In this paper both a particularly effective form of skeletal model called "s-reps" will be motivated and described, and methods of computationally deriving s-reps from object boundaries will be overviewed. In statistical applications, s-reps can provide an effective way to establish locational and orientational correspondences across the shape samples in the population. In this paper, not only the s-reps but also means of accomplishing the statistical objectives with respect to s-reps will be covered.

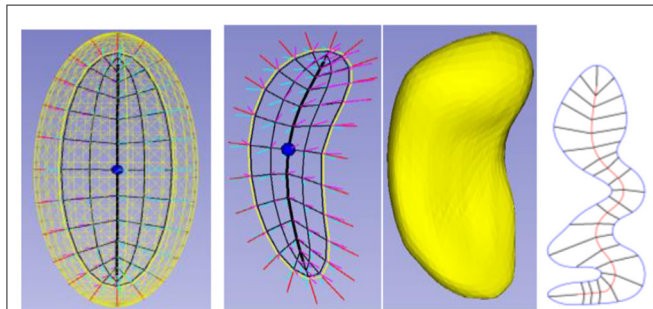
Many examples of applications of s-reps and complexes thereof, especially in medical-imaging-based data, will be given. However, the intent here is also to stimulate applications in many



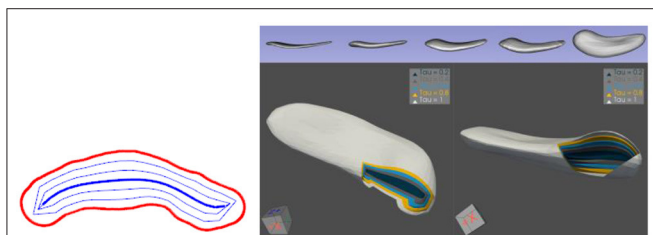
other areas of computer analysis and synthesis, including those on manufactured objects.

## SKELETAL MODELS AND S-REPS: DEFINITIONS AND MATHEMATICS

Roughly speaking, a skeletal model of an object (**Figures 1–3**) consists of a) a skeletal locus somehow central along the object and b) spokes, vectors emanating from the skeleton and ending at the object boundary, such that the spokes do not cross within the object. Traditionally, the skeleton has been considered a possibly branching surface or curve, but I find it much clearer to consider it as a collocated pair of surfaces or curves formed by a sort of collapse of the object boundary or better, which can be dilated to form the boundary implied by the skeleton and its spokes (**Figure 4**). In this way of thinking, the skeleton shares topology with the object boundary; it just has the constraint that except along the curve where the skeleton folds, the two copies of the skeletal surface or curve share the same skeletal spoke end point. This point of view allows the possibilities of having different spoke statistics for one of the skeletal copies than the other. Later in this paper, I will cover possibilities of relaxing the constraint of being a perfect skeletal copy in certain segments of the skeleton.



**FIGURE 3 |** Skeletal models. Left: Ellipsoid (3D) with its elliptical skeleton (grid), spine (bold), and center point (bullet). Middle left: A skeleton and its spokes for the hippocampus (3D) shown in middle right. Only samples of the spoke vectors are shown, but they exist continuously, i.e., with their tails at every continuous point of the skeletal surface. Right: A skeleton (red) and its spokes for a 2D object.



**FIGURE 4 |** Left: Radial onion skins in 2D. Right: Onion skins at  $\tau_2 = 0.0, 0.2, 0.4, 0.6, 0.8$ , and  $1.0$  (boundary) for a hippocampus; top row: individual onion skins for the various  $\tau_2$  values. Bottom row: two views of those onion skins seen end-on for a cut through the object.

The history and mathematics of skeletal models, as well as a number of algorithms for extracting them from object boundaries, is covered in detail in the book by Siddiqi (2008). Harry Blum, the inventor of the earliest form of skeletal model (Blum and Nagel, 1978), which we now call the “Blum medial axis,” felt that the major strength of medial models was that it provided a subdivision of an object into various attached parts. This goal turned out to be unachievable due the extreme bushiness: deep, broad, and random branching of the Blum medial axis; this bushiness was a consequence of the inevitable noise in the object boundary—this will be discussed at length shortly. However, the facts that the spoke lengths were a measure of object (half-)width and that the spoke directions and their derivatives were important indications of local object orientation and curvature have turned out to be particularly powerful measurements in object representation.

Blum, an inspired engineer, defined the Blum medial axis (that he called the “symmetric axis”) in terms of a flow at a constant rate (in the Euclidean metric in the object’s ambient space). He called this flow by the word “grassfire,” collapsing the object boundary to the medial axis. Blum and Nagel (1978) developed some early mathematics of this mapping from boundary to this skeletal structure, and many mathematicians took up the challenge of providing further properties; of special interest was the definition of a generalization they called the “symmetry set.” In it the spokes were always orthogonal to the object boundary, a property Damon calls “partial Blum.” Thus, it captured the boundary normals. The work of Srivastava et al. (2011) and others has emphasized that the behavior of these boundary normals is an important characteristic of shape. The way that they swing as you move along the boundary is the curvature information promulgated by Gauss (Koenderink, 1990).

The study of the symmetry set culminated in its singularity theoretic analysis by Giblin et al., a beautiful summary of which appears in Giblin and Kimia (2008). This work led to the important understanding that branching of the Blum skeleton was generic in both 2D and 3D.

The experience of hundreds of authors attempting to produce algorithms to map the object boundary to the Blum medial axis yielded an understanding that small protrusions or indentations in the boundary, either real or produced by noise, led to an extraordinary bushiness of branching and that the bush had to be heavily pruned if the skeletal structure was to be of any use—but especially in 3D robust algorithms to do this pruning were elusive. Certainly, little success in providing statistics on these derived structures was achieved due to the variation of this branching structure over a population of shapes.

The group I have led concluded that the branching structure needed to be fixed in order to support statistics. Supporting this need, my colleague James Damon invented a “reverse” type of flow, which he called “radial flow” (**Figure 4**), from the skeleton to a close approximation of the boundary. In it the skeleton, given its spokes, flowed (dilated) at a rate proportional to the spoke length. That is, for a spoke of length  $r$ , positions along the spoke were considered as  $\tau r$ , where  $\tau = 0.0$  on the skeleton,  $\tau = 1.0$  on the object boundary, and  $\tau = 0.5$  halfway from the spoke end on the skeleton and the spoke end on the boundary.

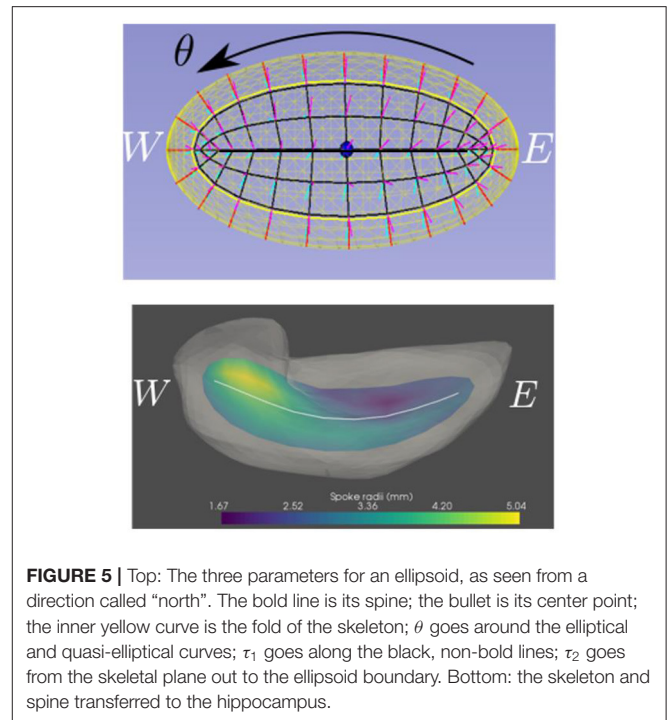
Damon (2008), and in more mathematical detail (Damon, 2003, 2004), discovered the mathematics of radial flow and the idea of the “onion-skin” surface at any given radial distance  $\tau$  from the skeleton; a particular accomplishment was his invention of a matrix-valued function on the skeleton,  $S_{\text{rad}}$  that yielded the velocity at which the spokes swung as you moved along the skeleton and a pair of eigenvalues called radial curvatures, whose comparisons to the corresponding boundary curvatures allow preventing spokes from crossing in the interior of the object. All this applied not only to the Blum skeleton but also to Damon’s generalization that did not need to meet all of the Blum properties. Among other things, the two spokes with collocated tails did not need to be of equal length and they did not need to be orthogonal to the boundary (be “partial Blum”). In order for the entity to capture width and direction properties, I added soft restrictions that this axis should rather closely achieve the properties of the Blum medial axis (Pizer et al., 2013; Liu et al., 2021a), and I named this form of skeletal structure the “s-rep.” S-reps are in this generalized skeletal category; almost always they are not Blum-medial and are designed to support statistical analysis.

The design was that a discretized form of the s-rep would be fit to the boundary, avoiding the problem of bushiness. As described and discussed in the next section, my colleagues and I produced software that by such fitting yielded the s-rep whose thus-dilated skeleton well fit the object boundary in both 2D and 3D and would with some tolerance meet the partial-Blum objectives.

For multiple objects, the shape information should include not only the shape of the individual objects but also their geometric relationships. Blum and Nagel had already described the “external medial axis,” which is the Blum medial axis of the complement of the objects. However, as pointed out by Damon and Gasparovic (2017), this axis was necessarily branching for 3 or more objects, even if the objects’ skeletons had consistent branching. But using radial flow past the objects’ boundaries, Damon extended the flow to what he called a linking axis, which typically branches, and as described in Section Multi-object Statistics Using Skeletal and Boundary Fitted Frames, Liu showed how to produce such a flow that does not branch.

After some experimentation with representations that included object angle, which were rejected because the bisector of the object angle was ambiguous when the angle was  $\pi/2$ , the UNC team settled on a representation by the spoke skeletal location, the spoke direction, and the spoke length (object’s local half-width) on the skeletal grid positions, since these emphasized the properties on which statistics could best be focused, we thought. Some applications represented a spoke by the coordinates of its two endpoints (skeletal and boundary).

I have come to see slabular objects as diffeomorphically deformed versions of the most basic slabular object, the (3D) ellipsoid. This view makes it analogous to the methods of shape representation via diffeomorphisms of a base object, but here the diffeomorphism is specifically on the object boundary and interior and we can take particular advantages of the inherent parameters of the skeletal positions for an ellipsoid (Figure 5) to produce correspondences needed for statistics.

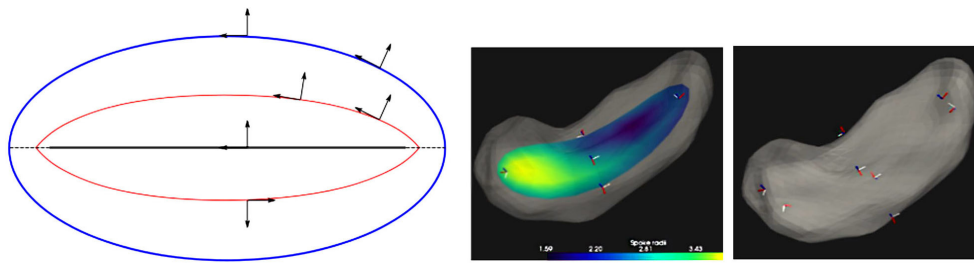


**FIGURE 5** | Top: The three parameters for an ellipsoid, as seen from a direction called “north”. The bold line is its spine; the bullet is its center point; the inner yellow curve is the fold of the skeleton;  $\theta$  goes around the elliptical and quasi-elliptical curves;  $\tau_1$  goes along the black, non-bold lines;  $\tau_2$  goes from the skeletal plane out to the ellipsoid boundary. Bottom: the skeleton and spine transferred to the hippocampus.

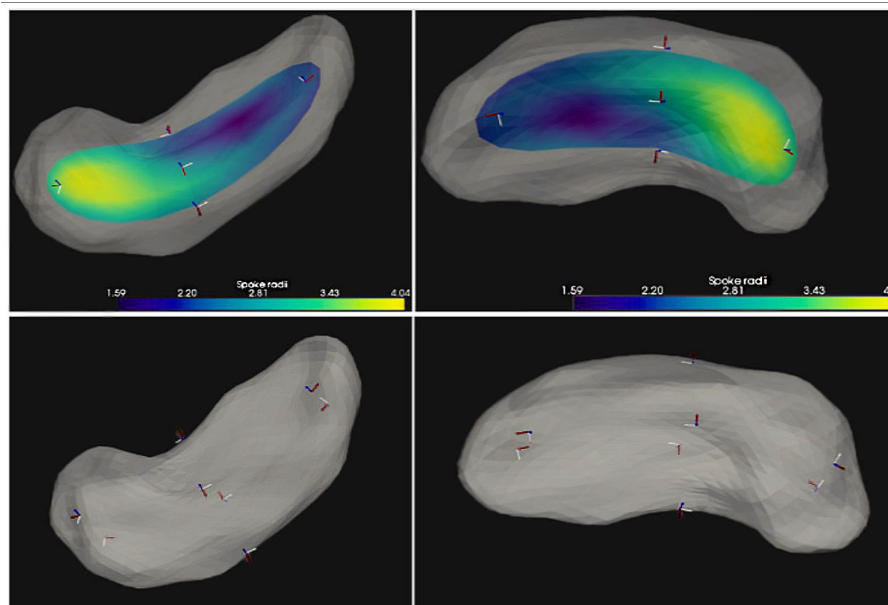
Let us first understand a 2D medial representation of an ellipse. That (2D) ellipse, which is the medial surface for an ellipsoid, can be represented by its skeleton, which we call the object’s “spine,” made from a (1D) folded line, parameterized within its interior by the cyclic value  $\theta \in [0, 2\pi]$ , together with spokes from that skeleton to the ellipse’s boundary.  $\theta$  is taken to be 0 at the center of the “north” side of the spine and to be  $\pi$  at the center of the “south” side of the spine. The spokes of the ellipse are parameterized by its radial flow value  $\tau_1$ , which, except for the spine ends ( $\theta = \pi/2$  or  $-\pi/2$ ), where  $\tau_1 \in [0, 1]$ , we take it to be in  $(-1, 1)$  with the sign indicating which side of the spine the spoke is and  $|\tau_1|$  indicating the radial distance from the spine. Moreover, the (1D) spine can be represented by its skeleton, a 0D entity, i.e., a point, which we call the object’s center point and choose it on the “north side” of the skeleton, i.e., to be at  $\theta = 0$ .

In fitting the s-rep to a new object by a diffeomorphism of the ellipsoid, including its boundary, its skeleton, its skeletal spokes, and its 3D spokes parameterized by the fraction multiples  $\tau_2$ , these ellipsoid points are carried by a diffeomorphism (warp) to form that new object, i.e., its skeletally implied boundary, its skeleton, its spine, and its center point. The skeletal coordinates of the ellipsoid are carried to their corresponding points in the target object; the desire that they are also radial lengths of the target object is so far not strictly met. This is discussed further in Section Fitting S-reps to Single- and Multi-object Boundaries.

As described shortly, a discretization of this representation is used in the SlicerSALT online toolkit. An interesting object type is a generalized cylinder, which is a representation by a central (skeletal) curve and its cross-sections. In that case one can think of the ellipsoid for which that object is a diffeomorphism



**FIGURE 6 |** Fitted frames at various places: Left: Frames fitted to skeletal and boundary points on an ellipse in 2D. The solid, thicker, black line is the skeleton, here the spine; the dashed lines are the skeleton extension; the red curve is an onion skin; the blue curve is the ellipse boundary; arrows are local frames. Middle and Right: The hippocampus (gray); middle: with its s-rep's skeleton (colored, with the color at a place on the skeleton showing the width of the object at that place) and fitted frames (blue element is normal) at the object center point and four points along and across the skeleton; right: the hippocampus boundary with fitted frames at points at ends of spokes at the skeletal points in the middle frame.



**FIGURE 7 |** Skeleton and fitted frames for a hippocampus from two points of view. Upper 2 panels: frames on the skeletal locus, colored by object width, where blue the object is thin, and where yellow it is thicker. Five fitted frames are shown on the skeleton, with the normal shown in blue; their orientation carries information as to the curvature of the skeleton. Lower two panels: the hippocampus boundary with fitted frames at points at ends of spokes at the skeletal points in the middle frame.

as being very eccentric, i.e., having one of its principal radii very much longer than the other two. In that situation the ordinary 3D skeletal analysis yields a spine, or its extension to the diffeomorphism of the endpoints of the ellipsoid, and this spine can be taken as the central curve. Moreover, the “cross-sections” are then defined by the skeleton parameterized by  $(\theta, \tau_1)$  and  $(\theta + \pi, \tau_1)$ , where  $\theta \in (-\pi/2, \pi/2)$ .

Repeating the idea of the transformation from a 3D ellipsoid to its medial representation of a 2D folded ellipse, and thence from one of the ellipses to its medial representation of a 1D folded line, results in the further transformation from one of the lines to its 0D medial representation, a center point. When the ellipsoid is carried to a target object by an appropriate diffeomorphism, that center point is carried to a place within the object that can be taken as its center. This location, being guaranteed to

be within and reasonably central in the object, is a far better representation than the object's center of mass, which can even be outside of the object. Finally, this idea can be generalized to higher dimensional hyperellipsoids with its principal radii notably different than the others and sortable into an increasing sequence. The succession of 1-lower dimensional hyperellipsoids, by a medial description of the just higher dimensional ellipsoids would allow our skeletal ideas to be generalized to dimensions higher than 3.

Realizing that rotationally and translationally normalizing a set of s-reps was a difficult challenge, Taheri and Schulz (2021) took Cartan's idea of representing space curves and surfaces using fitted frames and applied that idea to s-reps. Based on Taheri's inspiration, I created the following structure for a fitted frame (see **Figures 6, 7**) (Pizer et al.,



2021) that is consistent with skeletal geometry and with the fact that our s-reps are fit according to a diffeomorphism of an ellipsoid. It is based on the two radial flows just described, parameterized respectively by  $\tau_1$  and  $\tau_2$ . Thus, any point inside the object represented by the s-rep is determined by a  $(\theta, \tau_1, \tau_2)$  triple, each a function of Euclidean position within the object. If, as designed, common ellipsoid values of  $(\theta, \tau_1, \tau_2)$ , coming from diffeomorphisms from a common ellipsoid, yield statistically important correspondences, the fitted frames will yield statistically important orientation information.

The first vector of the fitted frame at any such point is taken to be tangent to the curve for the fixed  $(\tau_1, \tau_2)$  as the spine parameter  $\theta$  varies, a direction tangent to the 3D onion-skin at parameter value  $\tau_2$ . The second fitted frame vector there is the normal to the  $\tau_2$  onion-skin, with a sense away from the skeleton. Thus, on the spine of the skeleton ( $\tau_1 = 0, \tau_2 = 0$ ) the fitted frame has vectors along the spine and orthogonal to the skeletal surface. Off the spine for  $\tau_1 > 0$  but on the skeleton, the fitted frame has one vector along the skeletal curve for  $\theta$  varying and with fixed  $\tau_1$ , and a second vector orthogonal to the skeleton. On the boundary implied by the s-rep (for  $\tau_2 = 1.0$ ) at some  $\theta$  and  $\tau_1$ , the first vector in the fitted frame is tangent to the implied boundary there, and the second vector in the fitted frame is normal to the implied boundary there.

Now, following Elie Cartan's idea, geometric entities at a point should be understood according to the local fitted frame. The rotations of the fitted frames at any point interior to the object characterizes the local curvature of the object independent of rotations and translations of the object, which are important features. In **Figure 6**, for example, the rotation of the hippocampus skeleton from its center point to the end of the spine can be understood in terms of how the frame at the center point rotates into the frame at the end of the spine. Also, the rotation from the center point to its corresponding position on the boundary can be understood through the two frames at these positions. These rotations, as a function of three dimensions of motion, capture object curvature and can be fully characterized by three linear functions on a vector in 3-space (1-forms) measuring, respectively the rotations of the  $\tau_2$ -level surface normal ( $\nabla\tau_2$ ) into the other two respective frame vectors in the tangent space and the rotation of one of those  $\tau_2$ -level surface frame vectors into the other. Not only these curvatures but also all other s-rep-relevant vectors when expressed in that fitted frame at the tail of the vector, are invariant to rotations and translations of the objects. Examples of such features, are the tangent to the spine, the skeletal spokes of the spine, and the 3D spoke directions at each sampled spine position and the vectors from each sample point to its neighbors in the  $\nabla\tau_1$  (along the 3D spokes),  $\nabla\tau_2$  (from the spine toward the skeletal fold), and  $\nabla\theta$  (along the spine) directions. Damon has pointed out that the fitted frames and the associated features depend on the diffeomorphism used from the ellipsoid to the object and thus are not inherent unless the choice of diffeomorphism is in some sense inherent to the object (see Section Fitting S-reps to Single- and Multi-object Boundaries).

Nevertheless, as presented in Section Statistics on S-reps for Single and Multiple Objects, Taheri and Schulz (2021) and Liu et al. (2021b) have shown serious advantages to classification and hypothesis testing when discretized features according to the fitted frame were used in the statistics.

## DISCRETE S-REPS

There have been two general ways suggested for discretizing a skeletal model. Yushkevich et al. (2003) did that by computing an appropriate spline, which discretizes by the discrete set of basis function coefficients defining the spline. In the work of my group the discretization is performed on each of the parameters representing the s-rep:  $\theta$ ,  $\tau_1$ , and  $\tau_2$ , into integer submultiples of their range, e.g., for the examples in **Figures 2, 3, 5**,  $\tau_2$  as 0,  $\frac{1}{2}$ , and 1;  $\tau_1$  as  $-1, -\frac{1}{2}, 0, \frac{1}{2}, 1$  except at the spine ends, where its sampled values are 0,  $\frac{1}{2}, 1$ ; and  $\theta$  into the cyclic values  $-\pi/2$  for the east end of the spine, to  $\pi/2$  for the west end of the spine, in steps of  $\pi/10$  for the north side of the spine, and  $\pi/2$  to  $3\pi/2$  in steps of  $\pi/10$  for the south side of the spine. When the spokes are sampled only into 0 and 1, this produces a mesh of quadrilaterals on both the north and south sides of the skeleton, and corresponding meshes on the north and south sides of the object boundary (**Figure 3**, 2nd from left).

## MULTI-FIGURE S-REPS

I use the term "figure" to refer to a geometric entity with an unbranching skeleton and which persists across a population of objects. An example is the coronoid process in the mandible (**Figures 1, mandible and 10**). Traditionally, a population of objects in which there is a relatively sharp protrusion or indentation in similar positions along the boundary has been represented via the Blum medial axis (**Figures 8A,B**), i.e., with a branching skeleton. There are three difficulties with this representation: 1). It does not explicitly distinguish which two branches correspond to the host figure and which corresponds to the protrusion subfigure; 2). The host figure has a nonsmooth locus in its skeleton; 3). There is a long section of the skeleton of the protrusion branch that accounts for a very short part of the boundary. In regard to difficulty #1, for the 2D medial axis Katz and Pizer (2003) developed a method based on a model of human vision to distinguish the host skeleton from the subfigure skeleton, but this method is not mathematically characterized and has not been generalized to 3D. A method with the same objective in 3D has been presented in Reniers et al. (2008), albeit based on a different approach than Katz's.

Han et al. (2005) developed an alternative representation for the skeleton of a 3D object that overcomes the aforementioned difficulties. It is made from a host figure, a protrusion or indentation subfigure, and a description of the relation between the two. Its host skeleton is smooth and, intuitively speaking describes the host as if it did not have the protrusion or indentation, but it has a hole punched into its domain of  $(\theta, \tau_1)$  into which the subfigure skeleton is smoothly attached. Its subfigure is either additive, for a protrusion, or subtractive, for an indentation, and it has a skeleton that has a fold at



one end but which is truncated at the other end. Moreover, he devised a skeletal mechanism that attaches the truncated end of the subfigure skeleton smoothly into the hole in the host skeleton (Figures 8C–E). The attachment, which we call a “skirt,” represents the small part of the object boundary transitioning from the subfigure to the host. In the skirt the skeleton’s two sides separate from each other; as a result, the interior object region outside of the skirt is understood in the radial coordinates  $(\theta, \tau_1, \tau_2)$ , but the inside of the skirt contains no spokes and thus does not have radial coordinates. Despite this drawback, this context of s-reps with fitted frames began to be worked on. In that situation, the relation between the subfigure and the host figure can be expressed in terms of the fitted frames of the two figures’ s-reps. This setup has seemed attractive, but only recently has the continuation of this work on host figures and subfigures begun.

The surfaces of many manufactured objects have curves or points where the curvature is sharp: boxes are a good example. It would appear that there are natural skeletons of such objects, but the Blum skeleton will just not do. Far from the ends orthogonal to the longest axis of the box, the Blum skeleton stops being parallel to sides along the shortest axis and branches into long tracks accounting for the edge curves and corner points of the box. **What is needed is for the skeleton to run from end to end along the long and medium-length axis directions, and it must also have some sort of specialized protrusion marker to account for the edges and corners of the box. If such a description were to be invented, it would allow one to bring to bear all of the other beneficial aspects of shape analysis via skeletons.**

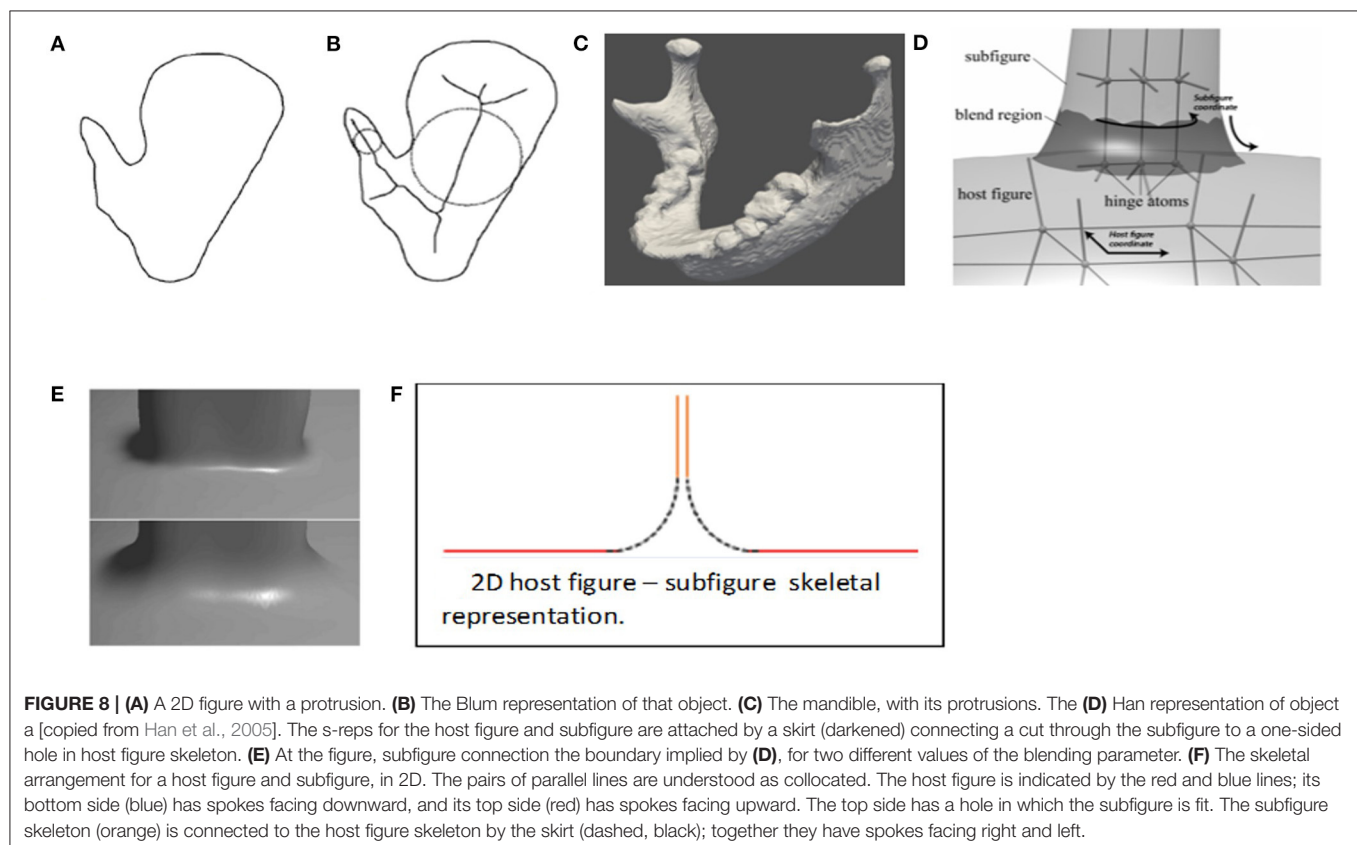
**Another problem with manufactured objects is that many, at least according to their design, have the form of the non-generic entities that have circular symmetry, for example wheels and balls. The skeletons of such objects have a different dimension than general s-reps, so special steps would need to be taken to include these representations.**

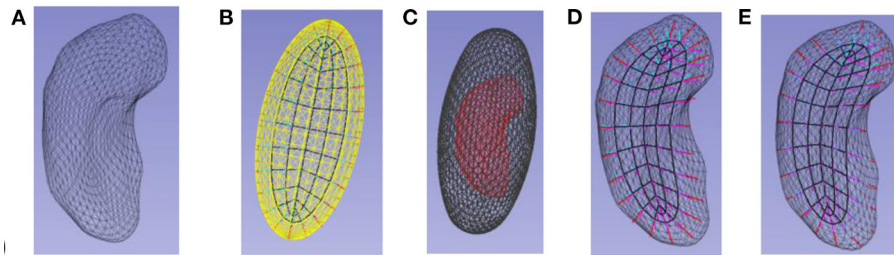
This concludes the description of the continuous s-rep. The following section describes the methods for fitting the discrete s-rep into an object described by its boundary.

## FITTING S-REPS TO SINGLE- AND MULTI-OBJECT BOUNDARIES

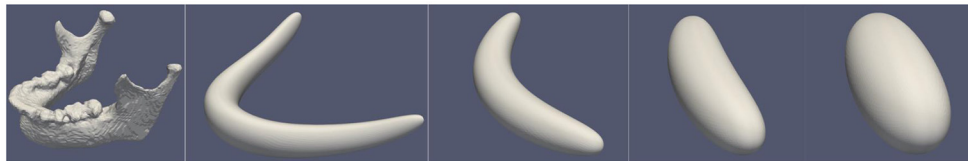
As introduced in Section Discrete S-reps, the discrete s-rep roughly is made from samples of the continuous s-rep, together with a mathematically appropriate means for interpolating the discrete s-rep into a continuous s-rep (Liu et al., 2021a). Typically, objects from images are represented by their boundaries or with binary images for which the cracks between within-object pixels or voxels and outside-of-object pixels or voxels form the object boundary. To utilize the benefits of s-reps, a means of fitting an s-rep to the object boundary is needed.

Why must the process of s-rep determination be a fitting to the input boundary and not a generation from the object boundary? Transformation of an object boundary to the skeleton has long been understood to be a process enhancing noise and detail. Despite hundreds of algorithms designed to accomplish this transformation, they have all foundered on the fact that little pimples or dimples in the boundary yield either a skeleton that





**FIGURE 9** | Fitting an s-rep into a hippocampus boundary. **(A)** The hippocampus, represented as a boundary mesh. **(B)** The ellipsoid, with its skeleton, into which **(C)** a hippocampus (red) is flowed. **(D)** The skeleton morphed into the hippocampus. **(E)** The refined s-rep.



**FIGURE 10** | Conformalized mean curvature flow on a mandible.

is bushily branching or one for which the implied boundary of the resulting s-rep is unfortunately far from the object boundary that was input to the process. Methods for pruning the bush have been developed in 2D (Ogneiewicz and Kübler, 1995), albeit with a result that suffers from the inaccuracy problem just mentioned. And in 3D, despite many attempts by strong scientists, no adequate pruning method has been found. The result is that s-reps produced by generation from the boundary are poorly suited to statistical applications. The variable branching is what causes the problem.

So instead of going from the boundary to the skeleton, with its major enhancement of noise or detail in the boundary, the inverse operation of going from an s-rep to the input boundary has been shown to be successful, at least for certain objects, because it accomplishes a smoothing of the noise. The basic idea, as I see it, is to diffeomorphically deform the simplest skeletally described object into the target object and to carry the skeleton of the simply described object into the target object via that diffeomorphism. The simplest skeletally described object is the ellipsoid in 3D and the ellipse in 2D. The challenge then becomes determining what the diffeomorphism should be.

The idea of fitting a skeleton (Liu et al., 2021a) to a target object boundary (see **Figure 9**) is to successively flow the boundary in an efficient way by smoothing it into an ellipsoid and then to reverse the flow while carrying the ellipsoid's s-rep with it. The initial try at this flow (Hong et al., 2016; Hong, 2018) moved the boundary along boundary normals at a rate monotonic with the curvature at each (so far smoothed) boundary point. When the flow rate is proportional to mean curvature, it is unstable at regions of high curvature, but Kazhdan et al.'s (2012) conformalized mean curvature flow solves this problem (**Figure 10**). For almost all slabular objects of interest, this flow approaches an ellipsoid (an ellipse if in 2D), and one can check whether the smoothed boundary output from some iteration is

close enough to an ellipsoid. The skeleton of that closest ellipsoid is analytically known. If that ellipse skeleton is deformed first to the approximate ellipse produced at the end of the flow, then the deformations provided by the smoothing iterations are successively reversed, and one has a diffeomorphism to the target object. This diffeomorphism can be applied to the skeletal and boundary ends of each of its discrete spokes, and the result is an s-rep for the target object. If the diffeomorphism does not correctly reflect skeletal properties (see item 1 in the next paragraph), the s-rep can be refined by a refinement diffeomorphism achieved by optimization of the following skeletal properties (Liu et al., 2021a):

- 1) foremost, a term heavily penalizing crossing of the spokes, via the comparison between boundary curvatures and the radial curvatures mentioned earlier;
- 2) a term penalizing the deviation of the implied boundary from the target object boundary;
- 3) a term penalizing the deviation of the angle of the spokes from the corresponding boundary normals.

**In addition, further closeness to mediality could be achieved by including a term penalizing the magnitude of the difference in the lengths of the two spokes emanating from the skeleton points that share a Euclidean location, and a penalty on the straightness of the spokes mapped from the ellipsoid.** The fitting algorithms found in the Slicer/SALT toolkit (Vicory et al., 2018) optimizes an objective function made from the first 3 terms, and at present it allows only the discrete spokes' lengths and positions on the boundary to move, but **it probably would be effective to allow their positions on the skeleton also to move.**

**Three things could be improved in this approach for producing geometric features in correspondence across a population:**

- 1) The fitted frames and their locations depend on the diffeomorphism from an ellipsoid to each object in the population. But no intrinsic, fully satisfactory way to produce diffeomorphisms that works across a wide range of object shapes and yields correspondence across a population of similar shapes has been produced. Indeed, the s-rep that we produce is not as close to skeletal as is desired, as it maps straight spokes into curved ones. This prevents the fitted frames from meeting their goals as well as they could and makes correspondence across objects in a population that is needed for statistics less strong than it needs to be. Rather than having a refinement step, it seems that a better idea is to require the stages of backward transformation to keep the spokes straight and the velocities along them fixed in radial distances. Damon (2021) has shown how to do this when transforming one ellipsoid to another. In ongoing work in our laboratory we are attempting to generalize this to any slabular object by insisting that the sequence of points along the spokes stay straight.
- 2) The method of fitted frames produces correspondence to the extent that objects in a population are geometrically similar and the diffeomorphism computation for each reflect that similarity. Thus, the fitted frames for these objects reflect that similarity. However, the method does not explicitly reflect the statistical variation within the population as such, nor does it reflect any biological correspondences.
- 3) Following the human vision property reported in (Burbeck et al., 1996) that skeletal properties are measured at a spatial scale proportional to the object width (spoke length), the fitting could use this multi-scale approach.

Fitting a skeleton in 2D (Figure 3, right) operates in essentially the same way as in 3D, except that the model is an ellipse (Krishna et al., 2022). The skeleton being formed from an ellipsoid's elliptical skeleton means that the curved skeleton formed by the diffeomorphism also has a spine and spokes, that is, the skeleton is also represented skeletally. Hong (2018) designed an extension of an s-rep for objects that can be understood to end in a cusp, such as the caudate nucleus. Vicory et al. (2022) has dealt with the problem of objects with multiple crests that must be put into biological correspondence by designing a diffeomorphism preceding the curvature-flow-based deformations, where the preceding flow smooths the high curvature regions in a way respecting their locations on the object within the members of its population. A principled method for this preliminary analysis, based on the shape statistics of the boundaries in the population from which a particular object is a statistical sample, would be useful.

## MULTI-OBJECT COMPLEXES

In many contexts, certainly in the human body, objects do not appear by themselves but in complexes of many objects. These objects can be separated or can share portions of their boundaries. In populations they can be adjacent or

can pull apart, and they can slide along each other. The shape of an individual object is often correlated with that of nearby objects.

In our laboratory multi-object complexes in populations began to be studied as diffeomorphisms in a space of many objects, but with one object's diffeomorphism related to its shape properties and also to neighboring objects' shape properties (Saboo, 2011). However, we came to realize that richer geometric properties than just voxel (or pixel) positions were needed to describe the necessary relationships.

Multi-object and multi-figure fitting benefits from a representation that captures not only each of the component objects or figures but also the relations between them. Damon and Gasparovic (2017) has considered extension of the objects' spokes into a "linking locus" where they meet, Liu et al. (2022) has implemented that scheme in a way that avoids folding. Krishna et al. (2022) considers objects that share a portion of their boundaries and uses a skeletal description of that shared region to capture inter-object information. This requires him to flatten that shared region before applying 2D s-rep fitting and then to restore the curvatures of the resulting 2D skeleton and spokes. He accomplishes that flattening by projection of the shared boundary region onto the skeleton of one (or both) of the objects and then projecting the skeleton back onto the ellipsoid's medial ellipse whence it came. Before that, he modifies the 3D spokes of the two objects to be collinear in the shared boundary region so that deformations will resist interpenetration or pulling away from adjacency of the two objects.

Taheri and Schulz (2021) describes a different means of providing a linking locus, in which the linking is first described among the ellipsoids from which the object were diffeomorphically deformed and then the linking is transformed by a diffeomorphism common to the objects.

An attractive property of these multi-object representations is that the fitted frame can be extended from the interiors of the objects to the space between the objects. The early stages of doing this have been accomplished by Liu et al. (2022).

These multi-object representations appear to be useful not only for studying inter-class relationships but also for segmentation and, especially, segmentation editing, as described in Section Statistics on S-reps for Single and Multiple Objects.

## STATISTICS ON S-REPS FOR SINGLE AND MULTIPLE OBJECTS

The correspondence between spatial samples across an object's training or testing population is critical to the success of the statistical operation. Because each s-rep for such a population is fit from a single shape and because the fitting is based on a rich set of shape features, s-reps do particularly well in producing correspondence. Tu et al. (2016) showed that even if the ultimate shape representation was a boundary PDM, choosing the spoke ends of a fitted s-rep as the boundary points yielded better statistical performance, at least on the objects she tested.

Because they have a heavy component of directional information and directions reside on unit spheres, the shape space on which s-reps reside is curved. One way to characterize an s-rep is as a tuple of  $n_4$  frames in 3D, a tuple of  $n_3$  3D directions, a tuple of  $n_2$  2D directions, a tuple of  $n_+$  positive variables such as lengths and an object size, and possibly a tuple of  $n_L$  3D locations. That is, an s-rep lives on  $(S^3)^{n_4} \times (S^2)^{n_3} \times (S^1)^{n_2} \times (R^+)^{n_+} \times (R^3)^{n_L}$ . As discussed in detail in Pizer and Marron, 2017, each of the positive variables can be mapped from  $R^+$  to  $R^1$  (Euclideanized) using the logarithm and then mean centered by subtracting  $\log(\text{the geometric mean of the variable})$ . The sphere-resident (directional) features need also to be Euclideanized before any of the standard statistical methods can be applied. We have accomplished those Euclideanizations using the Principal Nested Spheres method of Jung et al. (2012), which is a counterpart to Principal Component Analysis (PCA) for sphere-resident feature points. PNS is provided on the SlicerSALT toolkit (Vicory et al., 2018).

These representations for doing statistics require some sort of pre-alignment in position and orientation (and possibly spatial scale) to make the directions and locations consistent. Doing the alignment has the advantage that it can provide global object features for the statistics, such as its volume or the position of its center in a coordinate system based on some landmark. However, having to do the alignment generates difficulty because how to choose the alignment is unclear, even for a population of single one-figure objects. It is even more unclear for a population of multiple objects or multi-figure objects. The result is that the variability of the alignment adds noise to the representation and thus makes the statistical analysis less powerful.

As described in the following two subsections, Mohsen Taheri, Zhiyuan Liu, and Akash Krishna have used their s-rep fitted frames to generate s-rep features that do not depend on alignment. They have found these to provide particularly powerful hypothesis testing or classification on aspects of local geometry. This attractive idea has had only limited application because it is very recent, but I predict that it will become a method of choice quickly.

## Single Object Applications

Classification into s-rep classes simply uses s-rep features as the basis of classification, as initially studied in the work of Hong et al. (2016) on hippocampi between typical patients and first-episode schizophrenics and between hippocampi and caudate nuclei between typical 6-month olds and those who later developed symptoms of autism. He showed improvements in the classifications that used s-rep features over those that used only object boundary point features.

How many dimensions do the s-rep-based geometric properties involve? If the property is given by a fitted-frame or a rotation between a pair of frames, it lives on a hemisphere of a 3-dimensional sphere (understood to be embedded in 4 dimensions), so its representation requires a 3-tuple. Euclideanization of each of these thus yields 3 features. If the rotations of concern are with respect to positional changes

in all three basis directions, that information requires a 9-tuple when Euclideanized.

If the geometric property is a direction related to a particular position, such as a spoke direction in an s-rep, it can be understood to live on its own 2-dimensional sphere representing directions in terms of a relevant frame and thus to require a duple when Euclideanized. However, it may be useful to understand the direction with respect to more than one coordinate frame, e.g., the frame at the skeletal end of an s-rep spoke and the frame at the boundary end of the spoke. In that example, a 4-tuple would be needed to express the direction.

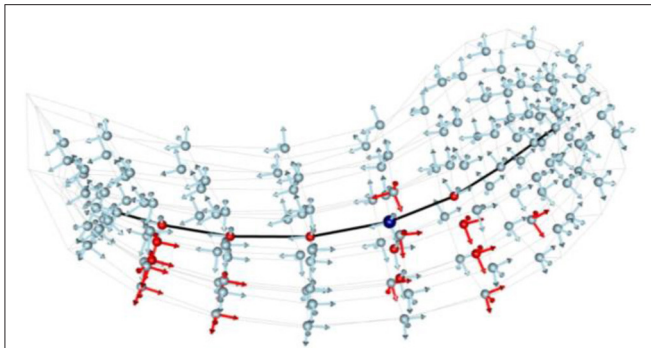
If the geometric property is a point's position, e.g., at some location in an s-rep's skeleton, that position needs to be understood in some frame, e.g., of the s-rep's center point, or even more than one frame. The position in each frame requires a 3-tuple. That 3-tuple can be expressed as a coordinate 3-tuple, or often more effectively as a direction (a duple) and a 1-dimensional, zero-mean Euclideanized length.

All of the aforementioned measurements can be made at many spatial scales to capture not only relations local to the s-rep but more global relations. After Euclideanization a PCA on the Euclideanized features is used to handle the covariance between the various curvatures, spokes, lengths, etc. Alternatively, as described in Sharma and Eltzner's research (Sharma et al., 2021), if the feature-tuples cluster on the Cartesian product of spheres, they can be mapped onto a high-dimensional sphere, on which Euclideanization can take place, directly capturing the covariance between the various direction vectors. However, research so far has shown little advantage to the alternative method of Euclideanization in various classification applications, so our standard technique is sphere-by-sphere Euclideanization.

Schulz et al. (2015) and Taheri and Schulz (2021) have used these features to do hypothesis testing between s-rep classes, studying which geometric object properties (feature tuples capturing a single geometric property, such as a spoke direction) differ significantly between the classes. Taheri showed in experiments comparing hippocampi between typical humans and those with Parkinson's disease that the fitted-frame based features produce superior detections of differences than those using global coordinates. The introduced features, including point locations and frame orientations, were measured based on each fitted frame for either an object or a complex, with Euclideanization for the orientations. Correcting for multiple tests was accomplished via the moderate approach of Benjamini and Hochberg (1995). Taheri and Schulz compared the left hippocampi of 182 patients with early Parkinson's disease (PD) vs. a healthy control group with 108 members. **Figure 11** illustrates the result of the tests where red spheres and red arrows indicate significant positions and orientations, respectively. They saw four significant point locations on the spine. One may conclude that there is no bending or twisting as orientations are similar on the spine. However, a concentration of significant point locations and orientations at the lower middle part of the hippocampi may reflect a bending.

The boundary ends of an s-rep's spokes imply a boundary Point Distribution Model (PDM). In a test data set of hippocampi Tu et al. (2016) showed that PDMs based on entropy-based





**FIGURE 11 |** Hypothesis tests on hippocampi of Parkinson's Disease vs. Control Group. Light-blue and red arrows indicate fitted frames representing non-significant and significant orientations, respectively. Light-blue and red spheres depict non-significant and significant positions. The dark-blue bullet is the center point. The black curve is the spine. Results are after  $p$ -value adjustment by Benjamini-Hochberg with False Discovery Rate (FDR) equal to 0.01.

correspondence of s-reps yielded better statistical properties than PDMs formed by entropy-based correspondence of the points themselves (Cates et al., 2006). Liu et al. (2021b) evaluated the jointly varying Euclideanized features from an s-rep implied multi-object PDM using the AJIVE method (Feng et al., 2018). He found that these jointly varying features produced better classifications of the hippocampus, caudate nucleus pair than the concatenation of the individual PDMs from the two objects.

Segmentation of objects from images requires knowledge both of the object shapes in the population and of the appearances in the image. Each of these can be represented by a probability density involving the object representation  $\mathbf{z}$ :  $p(\mathbf{z})$ , giving the shape information, and by the conditional density  $p(\mathbf{I} | \mathbf{z})$ , where the elements of  $\mathbf{I}$  are image intensity features, giving the appearance information. There has been quite a lot of research in our group on segmentations using s-rep features to make up  $\mathbf{z}$  and producing the segmentation as the most probable  $\mathbf{z}$  given  $\mathbf{I}$ . When Bayes theorem is applied, it follows that the  $\mathbf{z}$  resulting from this segmentation approach is  $\arg \max_{\mathbf{z}} [-\log p(\mathbf{z}) + (-\log p(\mathbf{I} | \mathbf{z}))]$ . This approach was most heavily developed in the segmentation of organs in the male pelvis from CT for planning of radiation therapy of prostate cancer (Levy et al., 2007), and the methods were the basis of segmentation by a spinoff corporation, Morphormics (Holloway et al., 2008), which was later bought by Accuray. Vicory (2016) applied this notion to the segmentation of the prostate from 3D ultrasound, given its shape in MRI. That is, the probability densities needed were on shape change, as opposed to on shape. This required a normalization of the MRI shape in both the training cases and the target cases before the statistics could be computed or used. This normalization was accomplished by applying a mean s-rep deformation to the s-rep from the patient's MRI before finding the shape change with the maximum posterior.

In Vicory's work the intensity features were not only image intensities but also derived texture features. Also, the appearance log probability was based on probability densities giving the

probability of a voxel being inside the object given the intensity and textures tuple.

Certain objects produced by automatic segmentation methods need editing via user interaction. When the editing is done for a 3D object, editing the image slice by slice is too time-consuming. Thus, segmentation editing is an important objective. We believe that this editing should combine some but limited user specification on image slices, geometric information from acceptable segmentations in image slices, and image appearance information. Mostapha et al. (2017) built a system based on s-rep statistics on shape changes needed, but it did not include a basis on appearance information. **Also needed would be s-rep-based shape changes of neighboring objects conditioned on the changes of the prime objects. That work is yet to be done.**

## Evaluations via Statistics

Since s-reps were designed to produce features useful for statistics on shape and in particular to produce spatial and orientational correspondences needed for statistics on populations of shapes, the evaluations need to be according to measurements of statistical success. The chapters (Pizer and Marron, 2017; Pizer et al., 2019) give many such measurements. In summary, on all of the anatomic objects tried and both of the diseases investigated for shape effects, the s-rep features and the methods of nonlinear statistics produced superior classification and hypothesis testing to the more traditional methods based on boundary points. Also, measures of generalization and specificity of the derived probability distributions led to preference for s-reps. Whether these preferences would follow through for features based on diffeomorphisms in a containing space has not yet been investigated, but the fact that object width features have been shown to be important for good statistics and these features are not directly available from diffeomorphisms suggest that even there statistics via s-reps might prevail.

The next section gives some recent results showing that statistics based on multi-object features also has strengths relative to alternative models.

## Multi-Object Statistics Using Skeletal and Boundary Fitted Frames

A particular capability of s-reps is how it enables powerful statistics on multiple objects. We present two approaches that have been developed.

### Classification and Hypothesis Testing on Separated Objects

Zhiyuan Liu has completed dissertation research (Liu, 2022) using data from two separated subcortical brain objects in infants, as imaged by MRI: the hippocampus and the caudate nucleus (Figure 12). The data fall into two classes related to whether the child will develop autism. He has developed multiple geometric features that can be used in classifications into the two classes, as well as hypothesis testing on the relation of features to classes. On these bases he has compared the use of various s-rep-related features. The results suggest that the following aspects are especially important:

Multi-object features that include between-object linking information, combined with within-object features classify more strongly than just within-object features.

- i. S-rep based features are particularly powerful, and especially affine frames produced by mapping s-rep fitted frames from the base ellipsoid to the target object.
- ii. A desired set of between-object features are lengths and directions of link vectors extending s-rep spokes in a 1-to-1 fashion to a linking surface between the objects.
- iii. Focusing on features varying jointly as an effect of a disease provides advantage over features selected without regard to joint variation.

These conclusions are also presented in two papers (Liu et al., 2021b, 2022). The first of these describes a method called NEUJIVE in which Euclideanized multi-object features are analyzed into joint features using the statistical method called AJIVE (Feng et al., 2018). The second describes how representing between-object shape using links from one object to a linking

surface between objects provides superior classification and hypothesis testing.

Here we briefly detail both the definitions and use of affine frames and of the linking vectors. The triplet of vectors in each affine frame are no longer unit length nor need they be orthogonal. They allow avoidance of preliminary alignment, which is especially challenging for multi-object complexes. They allow locations to be understood in the coordinate system of an object's skeletal center point. As well, the relation of each affine frame to that at that skeletal center point captures shape information itself.

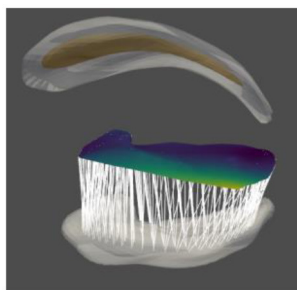
The linking surface is formed by a use of Damon (2021)'s linking mathematics. It smoothly and bijectively interpolates landmark pairs from equal length spoke extensions from the two objects' surfaces where there is no folding due to within-object spoke intersections. The links from an object to that surface are thereby smoothly interpolated to connect each discrete spoke to the linking surface (Figure 12), thereby depending on the good correspondence properties of discrete s-reps.

**However, what the best features are to describe inter-object relations is still an open question.**

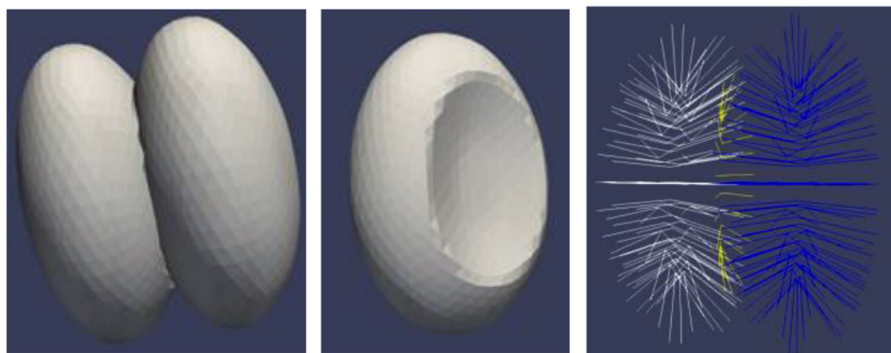
### Classification of Abutting Objects

Krishna et al. (2022) has work developing the ability to represent two objects with parts of their boundaries shared (Figure 13). His method involves providing s-reps of the two objects, where those s-reps' spokes are collinear within the shared boundary region. Moreover, he computes an s-rep of the 2D shared boundary as well. Not only are the features of the two objects understood in terms of their own fitted frames, but the s-reps features of the shared boundary region are understood in the fitted frames of one of the two objects.

Krishna compared classifications on deformed ellipsoid pairs sharing a boundary region. One experiment measured the accuracy of classifying a pair of deformed ellipsoids where one ellipsoid is bent from a pair of ellipsoids where neither object is bent. The other experiment measured the accuracy of classifying a pair of deformed ellipsoids where one ellipsoid is stretched



**FIGURE 12 |** An infant's hippocampus (surface shown at the bottom); the caudate nucleus, shown at the top by its boundary and skeleton; the inter-object linking surface in the middle colored by link length from the hippocampus; and the discrete links from the hippocampus shown for each discrete hippocampus spoke.



**FIGURE 13 |** Ellipsoidally based objects with a shared boundary that are used as data. Left: The two objects. Middle: one of the objects, showing the region of shared boundary. Right: the 2 object s-reps and the shared boundary s-rep (yellow).

relative to the other from a pair of ellipsoids where neither object is stretched. These classification capabilities were compared between s-rep features, using fitted frames, that ignored the shared boundary region's geometry vs. ones that included that geometry.

The number of s-rep points used was the same in all of the comparisons. His results show the benefits of inclusion of the shared boundary's s-rep features for object bending but not for object stretching. He also began work on a pair of brain structures that share part of their boundary. It did not adequately find a way to produce a smooth region of shared boundary from objects segmentations with individual segmentations in the form of coarse triangular tiles. **Future work should analyze brain structures having shared boundaries with such a technique. Also, a method using Liu's linking surfaces that include the shared boundary would be worthy of development.**

## DIFFICULTIES WITH AND LIMITATIONS OF SKELETAL REPRESENTATIONS

The skeleton of a 3D object is most well understood when the principal radii of the ellipsoid generating that skeleton are all notably different from each other. A particular problem is populations within which in one part of the population the longest object axis corresponds to the second longest axis in another part of the population or the population contains objects for which moving along the longest axis makes the second longer axis transition to be shorter than the one that was third longest—the transition is generic even though the transition shape, with a circular cross-section, is not. Moreover, when the smaller two of the principal radii remain close for an interval along the longest axis, the resulting near-circular symmetry makes the skeletal surface very thin and the orientation of the skeleton unstable. When this happens for objects considered as a quasi-tube, the skeletal orientation about the tubular axis (the spine) will seem to discontinuously change. **Work to deal with this behavior is needed.**

A strength of s-reps is that they are insensitive to noise in the boundary that yields pimples and dimples. Yet in some applications, e.g., where two objects must fit together tightly to form a seal against fluid leakage, the boundary must be expressed in a form that has no noise, e.g., with boundary intervals specified by splines. This is a real advantage of cm-reps (Yushkevich et al., 2015), where the implied or explicit spokes are normal to the boundary. **A challenge is to create a form of s-reps where subregions are restricted to having spokes normal to the boundary.**

While forms of fixed branching of at most a few levels could be easily handled by s-reps and their statistics, variable branching, such as happens in most tubular trees in the body, would require statistics of branching. Methods of the statistics of branching is a somewhat immature discipline, and its application to s-reps has not been accomplished.

In the simplest situations the fold of the skeleton in an s-rep is opposite a crest of the boundary. However, on the boundary's

crest region, along a principal curve crossing the crest, the zero level curve of the derivative of principal curvature can transition into an undulation with two crests and a trough. **In that case how the skeletal fold should behave has not been understood, to my knowledge.**

The major limitation of s-reps as the basis for statistics is that the data is typically provided as a mesh of triangular tiles representing the object boundary and that the s-rep must be fitted to that mesh. This weakness is shared with the many valuable methods for statistics based on computing a diffeomorphism over space including an object and then doing statistics on features derived from that diffeomorphism (see, for example, the Deformetrica library; Durrleman et al., 2014). An interesting alternative is the recently published work (Ambellan et al., 2021) that does its statistics directly on deformations of the mesh itself. **Research comparing these various methods for doing statistics on shapes would be valuable.**

Other limitations come from the fact that when corners and sharp edges are important features, as they are in manufactured objects, s-reps at present need to treat those somewhat unnaturally as subfigures. They also handle randomly branching objects such as blood vessel trees poorly, as well as objects in a single population that are in different topological classes or have different arrangements of their subfigures.

## DISCUSSION AND CONCLUSIONS

### Potential Nonmedical Applications

The chapter by Leymarie and Kimia (2008) lists a large number of applications to which skeletal models have been applied, from the cosmological scale to the atomic scale. **However, none of those used s-reps, but they could.** The following discusses applications where s-reps could have been used, and it discusses extensions of those as well as new opportunities for non-medical applications.

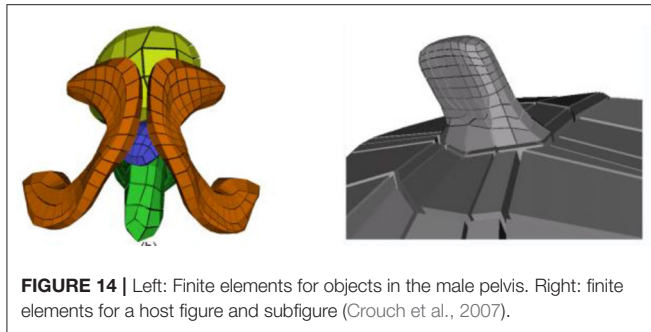
**Designing objects for manufacturing or art often nicely involves combining figures that can be understood skeletally. Means with an intuitive interface for specifying the shape of single figures using a limited set of primitives have already been worked on (Wikipedia<sup>3</sup>). However, s-reps would provide a much wider range of shapes as primitives. Means for controlling the way figures are pasted together would need to be invented, where these means were adequately intuitively controllable by the designer.**

**Computer graphics and the subset of that field toward visualization also could benefit from a wider set of primitives than are presently routinely used.**

The human body is made from articulated figures. Each figure, e.g., the forearm, has a recognizable shape, which can be statistically studied, not only statically but also in motion. While there are many studies of the body in motion as stick figures, it would seem helpful to study the body with flesh on its bones using the s-rep's efficient features. The ideas of figure-to-figure connections described above would appear to be useful here. Similarly, robots and other mechanical devices

<sup>3</sup>Wikipedia on 3D sculpting. Available online at: [https://en.wikipedia.org/wiki/Geometric\\_primitive](https://en.wikipedia.org/wiki/Geometric_primitive).





**FIGURE 14** | Left: Finite elements for objects in the male pelvis. Right: finite elements for a host figure and subfigure (Crouch et al., 2007).

made from articulated figures, such as airplanes, could benefit from using s-reps.

Once we have mechanical models formed from skeletons, it is natural to consider mechanical motion of these models. And this does not need to be restricted to articulation. Crouch et al. (2007) showed how skeletal models (albeit not s-reps) could be used to divide an object or a group of objects into natural elements for multi-scale finite element modeling of mechanical changes (Figure 14). For a single-figure object the subdivisions were along  $\tau_2$  and  $\theta$ . For multi-figure models including protrusions she showed how to have the finite elements transition from the subfigure elements to the host figure elements, providing an alternative to Han's approach described above. **Certainly, the shape properties revealed by a skeletal model make further work on physically based modeling using s-rep based elements an attractive direction.**

While s-reps have been used only for data that was extracted or is being extracted from medical images or images from ordinary cameras, its application to data extracted or being extracted from other sensors would be useful. For example, the LIDAR sensors used in self-driving vehicles would be an interesting source.

Human vision certainly divides the world into objects. Biederman (1987) and Burbeck et al. (1996) (in our group), and many others have adduced psychophysical evidence that it does so via skeletal primitives, which Biederman calls "geons." Burbeck and Pizer also gave evidence that the human visual system's skeletal analysis was done at spatial scales proportional to the object width. Indeed, there is some limited evidence that the monocular visual system is especially sensitive at skeletal points (Lee, 1995; Lee et al., 1998). Without some direct way for the brain to sense objects, how else could it be so fast in such sensing? Moreover, we have a good sense for categories of objects, e.g., faces and trees and roads. The instances within these classes differ geometrically from each other, but they typically have similar skeletal topology. **It seems natural to study mental models used in recognition by using s-reps.**

## Desirable Future Research on S-Rep Methodology

So far, when regions of the skeleton have been considered, they have been limited to ones entirely on either the north side or the south side of the skeleton. However, regions folded on a skeleton arise, for example, when an object abuts

another across a crest into which the skeleton fits. It should be straightforward to include such regions as a possibility.

So far, s-reps have been created only for single-figure or multi-figure slabular objects, i.e., those with spherical topology and have a skeletal surface, or for generalized cylinders, which also have spherical topology and focus especially on a curvilinear skeleton (Saboo, 2011). But there are many other topologies for which a skeletal model is appropriate. Cyclic forms of skeletal surfaces, such as closed fists, or of skeletal curves, for example, of doughnuts could be very useful. Likewise, s-reps for annular solids, such the myocardium would be useful.

In some populations a pair of objects in some instances share a boundary and in others the two objects are separated. It can even happen that one of the objects melds with the other object becoming a subfigure. Statistical methods for such populations could be developed. Also, one might want to be able to handle a host figure with two subfigures that can in some instances touch and in others be separated, and even meld together in such a way to change the number of holes (topological index).

Other situations needing development come when there are more than two objects. Handling how one object slides along the other two within the population could be handled via Liu's linking surfaces (Liu et al., 2022), which uses fitted frames. Also, the abutment arrangements can vary across cases, e.g., as one of the objects slides along the other. Cardiac valves can present such distributions.

Multi-scale s-reps would be worthy of study. For example, taking the spine as a whole at one spatial scale, without reflecting the shapes of the individual vertebrae, and then describing the vertebrae at a smaller scale, and the vertebral parts at a yet smaller scale would provide a driving problem.

Objects can be in motion. Especially when the motion involves deformation, as in the beating heart or breathing lung, statistical analysis of the motion sequences is of interest. Hong et al. (2019) has studied the progression of Huntington's disease statistically using cm-reps, Yushkevich et al. (2015) has studied objects in motion using his cm-reps, but such studies could usefully be extended to s-reps because the fitted frames will give powerful ways of characterizing the deformation. In general, comparisons between s-reps and cm-reps for a variety of applications would be informative.

It seems straightforward to apply this s-rep idea to higher-dimensional objects as long as they have codimension 1; i.e., where the ambient dimension is some  $n$  and the object boundary has dimension  $n-1$ . Far from straightforward, but likely important, would be the extension to a number of spatial dimensions and time. The difficulty is that a metric in (space, time) is complex because space and time are incommensurate. Making them commensurate would seem to require ideas of relativity. As exciting as this would be, it is beyond the scope of short term research, I believe.

Because the geometry of s-reps involves frames, which live abstractly in  $SO(3)$  (a hemisphere of  $S^3$ ), and directions, which live abstractly in  $S^2$ , the space describing a whole discrete s-rep lives on a Cartesian product of spheres, i.e., a polysphere:  $(R^+)^{d1} \times (S^2)^{d2} \times (S^3)^{d3}$ , where  $d2$  counts the number of vector



directions and  $d3$  counts the number of frame directions. The PNS algorithm for doing a PCA-like dimension reduction (Sharma et al., 2021) can be applied sphere by sphere, but it leaves the question of how to handle polyspheres that are more toroidal. That is, research on how from points on a toroidal polysphere of dimension  $d$  to create a subdimensional surface of dimension  $d-1$  that best fits the points is needed. Such research has recently been reported (Zoubouloglou, 2021) on Cartesian products of 1-dimensional spheres (circles), and **extension on the Cartesian product of 2-dimensional spheres is anticipated, but extensions to polyspheres made of 1-, 2-, and 3-dimensional spheres is needed.**

Good positional and orientational correspondence among instances of an object in a population is particularly important for statistical applications. Information as to this correspondence can come from understanding the objects in their source environment, e.g., biological correspondence for anatomical objects or structural correspondence for manufactured items. Or they can come from entropy analysis (Davies et al., 2001; Cates et al., 2006; Tu et al., 2016), though this can be very time-consuming and subject to local optima in optimization schemes. Or they can come from deep learning. **Methods for improving the correspondence by such means would probably be quite helpful to statistical or deep learning applications.**

**On the subject of deep learning for operations related to objects, such as recognition or segmentation, there is the open question of whether deep learners that are based on largely linear operators (other than the ReLus) on base elements such as voxel values or mesh node locations and links can compete with deep learners that use as features s-rep frames and connecting vectors, which are nonlinearly related to those more basic features.**

## CONCLUDING REMARKS

While not every object of study is suitable for representation by s-reps, there are many that are and that appear to be especially suitable for statistical analysis. This is because 1) object shape is best understood through a population of objects; 2) the s-rep methods presented here have notable advantages in providing descriptive object features with correspondence across the population, indicated by studies on anatomic objects. **While many of the theoretical challenges in developing and using s-reps have been met, there are many more remaining, and application opportunities abound.** The software in the SALT shape analysis toolkit (Vicory et al., 2018) supporting s-reps already can provide help to developments and uses of s-reps,

and more will be forthcoming from my laboratory. **Especially, bringing the advantages of s-reps to manufactured objects and to design remain an open challenge.** I hope that this paper will encourage others to take up this work. They will be welcome to add their software to the SALT s-reps collection.

## DATA AVAILABILITY STATEMENT

Inquiries on the data used in the many studies reported in this paper can be directed to the corresponding author. Much of this data is no longer available, and other of it is owned by other groups and is not publicly available.

## AUTHOR CONTRIBUTIONS

SP wrote the whole paper and solicited certain components from the other authors. ZL made many of the figures in the paper and wrote the program for fitting s-reps to boundaries that is reported in this paper. JM collaborated on much of the research reported in the paper and suggested improvements in the whole paper. JD collaborated on and created mathematics reported in the paper. JV collaborated on some of the research reported in the paper and provided a check on the reporting of that research in the paper. AK, ZL, and MT wrote the first drafts of the sections of the paper reporting their work, namely part of section 7.1 (MT) and section 2 (AK, ZL). All authors contributed to the article and approved the submitted version.

## FUNDING

This paper and its recent research was created with the partial support of NIH grant R01 EB021391, and NSF Grants IIS-1633074, and DMS-2113404. In the distant past the research was partially supported by other grants from NIH.

## ACKNOWLEDGMENTS

We are grateful to Nicholas Tapp-Hughes and Ankur Sharma for contributions to concepts and code for fitting s-reps to object boundaries. We are grateful to Drs. Martin Styner and Beatriz Paniagua for providing data for anatomic objects in populations. I am grateful to all of my former advisees who produced the many methods and applications described in this paper. I am grateful to Prof. Kaleem Siddiqi for collaborations in the area of skeletal models, for his co-authorship on our book (2008), and for his invitation as Associate Editor to write this paper.

## REFERENCES

- Ambellan, F., Zachow, S., and von Tycowicz, C. (2021). Rigid motion invariant statistical shape modeling based on discrete fundamental forms. *Med. Image Anal.* 73, 102178. doi: 10.1016/j.media.2021.102178
- Amenta, N., and Choi, S. (2008). "Voronoi methods for 3D medial axis approximation," in *Medial Representations*, Chapter 7 in eds K. Siddiqi, and S. M. Pizer (Springer).
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115
- Blum, H., and Nagel, R. N. (1978). Shape description using weighted symmetric axis features. *Pattern Recogn.* 10, 167–180. doi: 10.1016/0031-3203(78)90025-0

- Burbeck, C., Pizer, S. M., Morse, B. S., Arieli, D., Zauberman, G. S., and Rolland, J. (1996). Linking object boundaries at scale: a common mechanism for size and shape judgments. *Vision Res.* 36, 361–372. doi: 10.1016/0042-6989(95)00106-9
- Cates, J., Meyer, M., Fletecher, T., and Whitaker, R. (2006). “Entropy-based particle systems for shape correspondence,” in *1st MICCAI Workshop on Mathematical Foundations of Computational Anatomy: Geometrical, Statistical and Registration Methods for Modeling Biological Shape Variability*. 90–99.
- Crouch, J., Pizer, S. M., Chaney, E. L., Hu, Y., Mageras, G. S., and Zaider, M. (2007). Automated finite element analysis for deformable registration of prostate images. *IEEE Trans. Med. Imaging* 26, 1379–1390. doi: 10.1109/TMI.2007.898810
- Damon, J. N. (2003). Smoothness and geometry of boundaries associated to skeletal structures I: sufficient conditions for smoothness. *Annales Inst. Fourier* 53, 1001–1045. doi: 10.5802/aif.1997
- Damon, J. N. (2004). Smoothness and geometry of boundaries associated to skeletal structures II: geometry in the Blum case. *Compositio Math* 140, 1657–1674. doi: 10.1112/S0010437X04000570
- Damon, J. N. (2008). “Geometry and medial structure,” *Medial Representations*, eds Chapter 3 in K. Siddiqi and S. M. Pizer (Springer).
- Damon, J. N. (2021). *Thoughts on Ellipsoidal Models*. Personal Communication.
- Damon, J. N., and Gasparovic, E. (2017). Medial/skeletal linking structures for multi-region configurations, *Memoirs AMS* 250, 1–163. doi: 10.1090/memo/1193
- Davies, R. H., Coates, T. F., and Taylor, C. J. (2001). A minimum description length approach to statistical shape modeling. *Proc. Inform. Process. Medical Imaging* 2001, 50–63. doi: 10.1007/3-540-45729-1\_5
- Durrleman, S., Prastawa, M., Charon, N., Korenberg, J. R., Joshi, S., and Gerig, G. (2014). Morphometry of anatomic shape complexes with dense deformations and sparse parameters. *Neuroimage* 101, 35–49. doi: 10.1016/j.neuroimage.2014.06.043
- Feng, Q., Zhang, M., Hannig, J., and Marron, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.* 166, 241–265. doi: 10.1016/j.jmva.2018.03.008
- Giblin, P., and Kimia, B. B. (2008). “Local forms and transitions of the medial axis,” in *Medial Representations*, eds Chapter 2 in K. Siddiqi and S. M. Pizer (Springer).
- Han, Q., Pizer, S. M., Merck, D., Joshi, S., and Jeong, J. Y. (2005). “Multi-figure anatomical objects for shape statistics. *Inf. Process. Med. Imaging* 3565, 701–712. doi: 10.1007/11505730\_58
- Holloway, R., Pizer, S. M., and Broadhurst, R. E. (2008). *Autosegmentation of the Rectum. Video Published on the Web*. Durham, NC: Morphormics Inc.
- Hong, J. P. (2018). *Classification of neuroanatomical structures based in non-Euclidean geometric object properties* (Ph.D. dissertation). Department of Computer Science, University of North Carolina.
- Hong, J. P., Vicory, J., Schulz, J., Styner, M., Marron, J. S., and Pizer, S. M. (2016). Non-Euclidean classification of medically imaged objects via s-reps. *Med. Image Anal.* 31, 37–45. doi: 10.1016/j.media.2016.01.007
- Hong, S., Fishbaugh, J., Wolff, J. J., Styner, M. A., Gerig, G., the IBIS Network. (2019). Hierarchical multi-geodesic model for longitudinal analysis of temporal trajectories of anatomical shape and covariates. *Proc. MICCAI* 57–65. doi: 10.1007/978-3-030-32251-9\_7
- Jung, S., Dryden, I. L., and Marron, J. S. (2012). Analysis of principal nested spheres. *Biometrika* 99, 551–568. doi: 10.1093/biomet/ass022
- Katz, R., and Pizer, S. M. (2003). Untangling the Blum medial axis transform. In: *International Journal of Computer Vision - Special UNC-MIDAG issue*, eds O. Faugeras, K. Ikeuchi, and J. Ponce (Kluwer Academic), 55, 139–153. doi: 10.1023/A:1026183017197
- Kazhdan, M., and Solomon, J., Ben-Chen, M. (2012). Can mean-curvature flow be modified to be non-singular? *Comput. Graphics Forum* 31, 1745–1754. doi: 10.1111/j.1467-8659.2012.03179.x
- Koenderink, J. J. (1990). *Solid Shape*. Cambridge, MA: MIT Press.
- Krishna, A., Liu, Z., and Pizer, S. M. (2022). *Incorporating the Geometric Relationship of Adjacent Objects in Multi-Object Shape Analysis*. Internal report, Univ. of NC Dept. of Computer Science, under review for journal publication. Available by request from SM Pizer.
- Lee, T. S. (1995). “Neurophysiological evidence for image segmentation and medial axis computation in V1,” in *Fourth Annual Computational Neuroscience Meeting* (London; Academic Press), 373–378.
- Lee, T. S., Mumford, D., Romero, R., and Lamme, V. A. F. (1998). The role of the primary visual cortex in higher level vision. *Vision Res.* 38, 2429–2454. doi: 10.1016/S0042-6989(97)00464-1
- Levy, J. H., Broadhurst, R. E., Jeong, J., Liu, X., Stough, J., and Tracton, G. S. (2007). “Prostate and bladder segmentation using a statistically trainable model,” in *Published as Abstract and poster at conference of the American Society for Therapeutic Radiology and Oncology* (Arlington, VI: ASTRO).
- Leymarie, F., and Kimia, B. B. (2008). “From the infinitely large to the infinitely small,” in *Medial Representations*, Chapter 11 in eds K. Siddiqi and S. M. Pizer (Springer).
- Liu, Z. (2022). *Geometric and Statistical Models for Multi-object Shape Analysis* (Ph.D. dissertation). Department of Computer Science, University of North Carolina at Chapel Hill.
- Liu, Z., Damon, J. N., Marron, J. S., and Pizer, S. M. (2022). *Geometric and Statistical Models for Analysis of Two-Object Complexes*. Under review for journal publication.
- Liu, Z., Hong, J., Vicory, J., Damon, J. N., and Pizer, S. M. (2021a). Fitting unbranching skeletal structures to objects. *Med. Image Anal.* 70, 102020. doi: 10.1016/j.media.2021.102020
- Liu, Z., Schulz, J., Taheri, M., Styner, M., Damon, J., Pizer, S., et al. (2021b). Analysis of joint shape variation from multi-object complexes.
- Mostapha, M., Vicory, J., Styner, M., and Pizer, S. (2017). A segmentation editing framework based on shape change statistics. *SPIE Med. Imaging* 10133, 101331E. doi: 10.1117/12.2250023
- Ogneiewicz, R., and Kübler, O. (1995). Hierarchic Voronoi skeletons. *Pattern Recogn.* 28, 343–359. doi: 10.1016/0031-3203(94)00105-U
- Pizer, S., Krishna, A., Liu, Z., and Taheri, M. (2021). Fitted frames to object interiors.
- Pizer, S. M., Hong, J., Vicory, J., Liu, Z., Marron, J. S., Choi, H. -Y. et al. (2019). Object shape representation via skeletal models (s-reps) and statistical analysis. *Riemannian Geometr. Stat. Med. Image Anal.* 2020, 233–272. doi: 10.1016/B978-0-12-814725-2.00014-5
- Pizer, S. M., Jung, S., Goswami, D., Vicory, J., Zhao, X., and Chaudhuri, R. (2013). “Nested sphere statistics of skeletal models,” *Innovations for Shape Analysis: Models and Algorithms*, eds M. Breuss, A. Bruckstein, and P. Maragos (Berlin; Heidelberg: Springer), 93–115.
- Pizer, S. M., and Marron, J. S. (2017). “Object statistics on curved manifolds,” in *Statistical Shape and Deformation Analysis*, Chapter 6 in eds G. Zheng, S. Li, and G. Székely (Academic Press), 137–164.
- Reniers, D., Jalba, A., and Telea, A. (2008). “Robust classification and analysis of anatomical surfaces using 3D skeletons,” in *Eurographics Workshop on Visual Computing for Biomedicine*, eds D. Reniers, A. Jalba, and A. Telea (The Eurographics Association). doi: 10.2312/VCBM/VCBM08/061-068
- Saboo, R. (2011). *Atlas Diffeomorphisms via Object Models* (Ph.D.) dissertation, Dept. of Computer Science, Univ. of North Carolina.
- Schulz, J., Pizer, S. M., Marron, J. S., and Godtliebsen, F. (2015). Nonlinear hypothesis testing of geometric object properties of shapes applied to hippocampi. *J. Math. Imaging Vision* 54, 15–34. doi: 10.1007/s10851-015-0587-7
- Sharma, A., Eltzner, B., Huckemann, S., Marron, J. S., and Pizer, S. M. (2021). Comparison of two methods for Euclideanization in discrimination of geometric data on polyspheres. In preparation at Univ. of NC Dept. of Computer Science.
- Siddiqi, K. (2008). *Medial Representations*. Springer.
- Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2011). Shape analysis of rigid bodies in Euclidean spaces. *IEEE Trans. Pattern Anal. Mac. Intell.* 33, 1415–1428. doi: 10.1109/TPAMI.2010.184
- Székely, G. (2008). “Voronoi skeletons,” in *Chapter 6 in Medial Representations* (Springer).
- Taheri, M., and Schulz, J. (2021). *Statistical Analysis of Locally Parameterized Shapes*. Available online at: <http://arxiv.org/abs/2109.03027>.
- Tu, L., Vicory, J., Elhabian, S., Paniagua, B., Prieto, J. C., and Damon, J. N. (2016). Entropy-based correspondence improvement of interpolated skeletal models. *Comput. Vision Image Understand.* 151, 72–79. doi: 10.1016/j.cviu.2015.11.002
- Vicory, J. (2016). *Shape Deformation Statistics and Regional Texture-based Appearance Models for Segmentation* (Ph.D. dissertation). Department of Computer Science, University of North Carolina.

- Vicory, J., Herz, C., Han, Y., Allemang, D., Flynn, M., and Cianciulli, A. (2022). Skeletal model-based analysis of the tricuspid valve in hypoplastic left heart syndrome.
- Vicory, J., Pascal, L., Hernandez, P., Fishbaugh, J., Prieto, J., and Mostapha, M. (2018). "Slicersalt: shape analysis toolbox," in *Proceedings International Workshop on Shape in Medical Imaging* (Springer), 65–72. Available online at: <https://github.com/KitwareMedical/SlicerSkeletalRepresentation>.
- Yushkevich, P., Fletcher, P. T., Joshi, S., Thall, A., and Pizer, S. M. (2003). Continuous medial representations for geometric object modeling in 2D and 3D. *Image Vision Comput. Special Issue Generat. Modelbased Vision* 21, 17–27. doi: 10.1016/S0262-8856(02)00135-X
- Yushkevich, P., Pouch, A. M., Tian, S., Takebe, M., Yuan, J., Gorman Jr, R., et al. (2015). Medially constrained deformable modeling for segmentation of branching medial structures: application to aortic valve segmentation and morphometry. *Med. image Anal.* 26, 217–231. doi: 10.1016/j.media.2015.09.003
- Zoubouloulou, P. (2021). *Scaled Torus Principal Component Analysis*. Chapel Hill, NC: UNC.

**Conflict of Interest:** JV was employed by Kitware Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Pizer, Marron, Damon, Vicory, Krishna, Liu and Taheri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## OPEN ACCESS

## EDITED BY

Yi-Zhe Song,  
University of Surrey, United Kingdom

## REVIEWED BY

Shancheng Zhao,  
Jinan University, China  
Ruoyi Du,  
Beijing University of Posts and  
Telecommunications (BUPT), China

## \*CORRESPONDENCE

Jon Sparring  
sparring@di.ku.dk

## SPECIALTY SECTION

This article was submitted to  
Computer Vision,  
a section of the journal  
Frontiers in Computer Science

RECEIVED 09 May 2022

ACCEPTED 19 October 2022

PUBLISHED 17 November 2022

## CITATION

Sparring J and Darkner S (2022) An  
algebra for local histograms.  
*Front. Comput. Sci.* 4:939563.  
doi: 10.3389/fcomp.2022.939563

## COPYRIGHT

© 2022 Sparring and Darkner. This is  
an open-access article distributed  
under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#).  
The use, distribution or reproduction  
in other forums is permitted, provided  
the original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which  
does not comply with these terms.

# An algebra for local histograms

Jon Sparring<sup>1,2\*</sup> and Sune Darkner<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Copenhagen, Copenhagen, Denmark, <sup>2</sup>Center for Quantifying Images From MAXIV (QIM), Lyngby, Denmark

In this article, we consider local overlapping histograms of functions between discrete domains and codomains. We develop a simple algebra for local histograms. Based on a separation of overlapping domains into non-overlapping domains, we (1) show how these can be used to enumerate the size of the set of possible histograms given the local histogram domains, and (2) enumerate the number of functions, which share a specific choice of a set of local histograms. Finally, we present a decoding algorithm, which given a set of overlapping histograms, and calculate the set of functions, which share these histograms.

## KEYWORDS

infinitely-additive set functions, multisets, counting histograms and functions, locally orderless histograms, reconstruction from histograms

## 1. Introduction

Inspired by Koenderink and Doorn (1999), we have for many years worked with images, and features derived from local histograms, and a nagging question has been, what the degrees of freedoms remain, given a set of overlapping histograms. This paper presents a theoretical investigation into the relationship between sets of local histograms and functions between discrete domains and codomains of any dimension. We describe an algebra of histograms, which is strongly related to the algebra of sets on the function domain and multisets: Given a set of local histogram's domains,  $h(X_i)$ ,  $X_i \subset X$ , where  $X$  is the full domain, and  $X_i$  are subsets thereof, such that  $\bigcup_i X_i = X$ , we factor  $X$  into a new set of disjoint subsets  $\{X'_j\}$ ,  $\bigcup_j X'_j = X$ , and with this, we are able to count the number of independent histograms, which jointly describe the total set of local histograms, and which leads to a simple countable, generative model for functions drawn from these histograms. Finally, we present a simple algorithm for generating the set of functions, which share a particular set of local histograms overlapping or not.

Our work is an extension of Sparring and Darkner (2022), where 1-dimensional signals are considered and the concept of metameric classes is introduced in the concept of local histograms. The article restricts itself to binary signals from their densely overlapping histograms. In Wu et al. (2000), the authors consider normalized histogram of images filtered with Gabor kernels (Gabor, 1946), and in particular, the limiting case of the discrete domain converging to  $\mathbb{Z}^2$ .

This paper is organized as follows. In Section 2, we present the histogram-algebra, in Section 3 we show how the number of unique functions sharing a specific set of histograms is generated. In Section 4, we present the algorithm for calculating the set of functions, which share a given set of local histograms, and finally, Section 6 gives concluding remarks.



## 2. Histograms as Infinitely-Additive set functions

In the following, we will define an algebra for discrete histograms of disjoint domains, and we will extend this to non-disjoint domains by repartitioning domains.

Consider discrete domain  $X$ , co-domain  $A$ , and a functions  $f: X \rightarrow A$  between them, such that the histogram  $h: A \rightarrow \mathbb{Z}_+$

$$h_X(a) = \sum_{x \in X} \delta(f(x) - a), \quad (1a)$$

$$\delta(x) = \begin{cases} 1, & \text{when } x = 0, \\ 0, & \text{otherwise,} \end{cases} \quad (1b)$$

is defined. Conceptually, we think of  $X$  as  $d$ -dimensional spatial domain  $X = \{1, 2, 3, \dots, n\}^d$  with side-lengths  $n > 0$ , and  $A$  as an alphabet of  $m > 0$  different gray values  $A = \{1, \dots, m\}$ , but for the properties of possibly overlapping histograms, the interpretation of the values of  $X$  and  $A$  is not important, and  $X$  and  $A$  could as well be the set {cow, cat, fish} or color triplets  $\{(0, 0, 0), (0, 0, 1), \dots\}$ . As long as we can define a one-to-one mapping to an index set, we need only to concern ourselves with this index.

Two key properties of a histogram are that

**Property 2.1.** Histograms are non-negative,  $\forall a \in A, h(a) \geq 0$ .

**Property 2.2.** Every value  $f(x), x \in X$  is counted once and only once.

A direct consequence of Property 2.2 is that

$$\sum_{a \in A} h_X(a) = |X|. \quad (2)$$

In this article, we are interested in counting possible histograms and for given histograms, counting the number of possible function. Let's start by examining the number of unique histograms that exists for a single domain and co-domain. Let  $\mathcal{H}_X = \{h_X^i\}$ ,  $\forall i, j, h_X^i \neq h_X^j$  be the set of unique histograms. Its size may be calculated as unordered sampling with replacement, where we visually represent each element in  $X$  with a “•” and each bin edge with a “;.” Then the string “•••••;•;•••” corresponds to the histogram  $[1; 2; 3; \dots] \rightarrow [3; 1; 0; \dots]$ . For brevity, it is convenient to assume that an ordering of the alphabet exists such that we may write the before mentioned histogram simply as  $[3; 1; 0; \dots]$ . The string will be  $|X| + |A| - 1$  long, and all possible histograms can be produced by selecting  $|A| - 1$  positions in this string for the “;” character. Thus, the number of unique histograms for a given domain  $X$  is given by the binomial coefficient,

$$|\mathcal{H}_X| = \binom{|X| + |A| - 1}{|A| - 1} = \binom{|X| + |A| - 1}{|X|}. \quad (3)$$

In the following, we will consider possibly overlapping, local histograms over the domain  $X$ . Our expositions will be divided into first non-overlapping or disjoint domains, and then we will show how overlapping domains can be repartitioned into disjoint domains, and how these relate to the original overlapping domains.

### 2.1. Histograms over disjoint domains

Consider a partitioning of  $X$  into  $k < \infty$  disjoint subdomains  $X = \bigcup_{i=1}^k X_i$ , where  $\forall i \neq j, X_i \cap X_j = \emptyset$ . Due to Property 2.2,  $h$  is a finitely-additive set function (Stover, 2022), and hence,

$$\sum_{i=1}^k h_{X_i}(a) = h_{\bigcup_i X_i}(a) = h_X(a). \quad (4)$$

As a consequence,  $h_\emptyset(a) = 0$ , and addition of histograms of disjoint domains is commutative and associative. The subtraction  $h_Y(a) - h_X(a)$  is a histogram when  $X \subseteq Y$ , e.g.,

$$h_{X \cup Y} = h_X + h_Y \Leftrightarrow h_{X \cup Y} - h_X = h_Y \Leftrightarrow h_{X \cup Y} - h_Y = h_X, \quad (5)$$

omitting the argument  $a$  for brevity. However, subtracting any two histograms in general will likely produce negative values violating Property 2.1, and although useful at times, the result will not be a histogram.

Since the sets  $X_i$  are disjoint, the size of the set of all possible histograms of  $X$  is found by extending Equation (3) directly,

$$|\mathcal{H}| = \prod_{i=1}^k |\mathcal{H}_{X_i}|. \quad (6)$$

### 2.2. Partitioning of non-disjoint sets

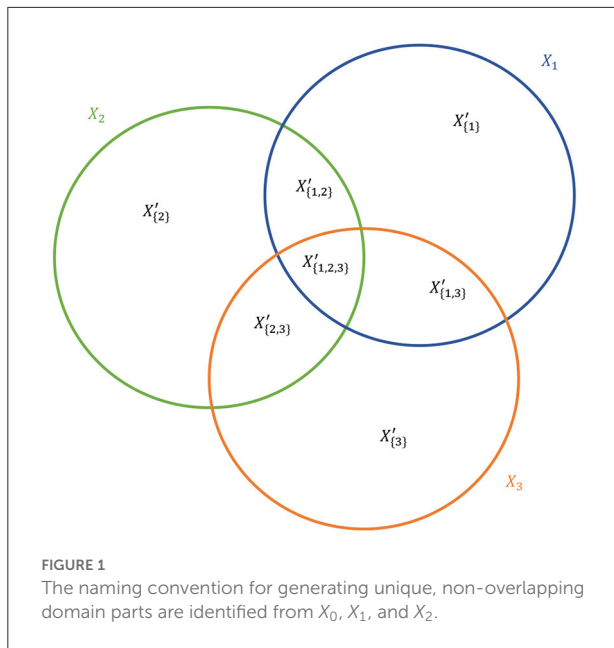
For a set of  $k$  non-disjoint domains  $\mathcal{X} = \{X_i\}$  of  $X = \bigcup_{i=1}^k X_i$ , we can repartition  $X$  into disjoint domains of unique overlap of  $X_i$

$$X'_I = \left( \bigcap_{j \in I} X_j \right) \setminus \left( \bigcup_{j \in \{0, 1, \dots, n-1\} \setminus I} X_j \right), \quad I \in P_k, \quad (7)$$

where  $P_k$  is the powerset of  $\{1, 2, \dots, k\}$ , e.g.,  $P_3 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$ . With this notation, we find the original sets as,

$$X_i = \bigcup_{p \in P_k : i \in p} X'_p. \quad (8)$$

**Example 2.1.** As an example, consider 3 sets  $X_1, X_2$ , and  $X_3$ , there are 7 unique intersections as illustrated in Figure 1 together with the powerset naming convention.



That is,  $X'_{1,2} = (X_1 \cap X_2) \setminus X_3$  and  $X_1 = X'_{1,1} \cup X'_{1,2} \cup X'_{1,2,3} \cup X'_{1,3}$ .

**Example 2.2.** As a concrete example, consider the domain  $X = \{1, 2, \dots, 6\}$  and the codomain  $A = \{1, 2, 3\}$ , and define  $X_1 = \{1, 2, 3, 4\}$ ,  $X_2 = \{3, 4, 5, 6\}$ . Assuming the usual ordering of integers, we can illustrate this overlap on a line as,

$$X = \underbrace{[1; 2; 3; 4]}_{X_1} \underbrace{[4; 5; 6]}_{X_2}. \quad (9)$$

Using Equation (7) we find that  $P_2 = \{\{1\}, \{2\}, \{1, 2\}\}$  and that  $X'_{1,1} = \{1, 2\}$ ,  $X'_{1,2} = \{3, 4\}$ , and  $X'_{2,2} = \{5, 6\}$ . Each of these subdomains are of size 2, and thus, the by Equation (3), number of possible histograms of each is  $\binom{2+3-1}{3-1} = 6$ , and the set of possible histograms is

$$\mathcal{H}_{X_I} = \{[2; 0; 0], [1; 1; 0], [1; 0; 1], [0; 2; 0], [0; 1; 1], [0; 0; 2]\}, \quad I \in P_2. \quad (10)$$

Since there are 3 disjoint regions each with 6 possible histograms, there are  $6^3 = 216$  combinations of these. Introducing a natural extension of our notations on the domains to their corresponding histograms, one of these is,

$$h_{\{1\}} = h_{X'_{1,1}} = [1; 1; 0], \quad (11a)$$

$$h_{\{1,2\}} = h_{X'_{1,2}} = [1; 0; 1], \quad (11b)$$

$$h_{\{2\}} = h_{X'_{2,2}} = [0; 2; 0], \quad (11c)$$

in which case,

$$h_1 = h_{X_1} = h_{\{1\}} + h_{\{1,2\}} = [1; 1; 0] + [1; 0; 1] = [2; 1; 1] \quad (12a)$$

$$h_2 = h_{X_2} = h_{\{1,2\}} + h_{\{2\}} = [1; 0; 1] + [0; 2; 0] = [1; 2; 1]. \quad (12b)$$

Since these overlapping histograms have been generated by histograms on their disjoint parts, we are sure that a function exists on  $X$  which has histograms  $h_1$  and  $h_2$ . Further, since histograms are finitely-additive functions we are sure that Properties 2.1 and 2.2 are fulfilled for  $h_1$  and  $h_2$ .

In the following, we will count the number of functions on disjoint domains and see how these can be combined to generate the family of functions, which share overlapping histograms generated from the disjoint domains.

### 3. Unique functions and their histograms on disjoint domains

For a single domain  $X$ , the total number of possible functions is given as  $|A|^{|X|}$ , and some of these have the same histogram. Conversely, given a histogram  $h$ , the set of functions, which share this histogram can be produced as the set of distinct permutations of the function,

$$S = \underbrace{[1; \dots; 1]}_{h(1)} \underbrace{[2; \dots; 2]}_{h(2)} \underbrace{[3; \dots; 3]}_{h(3)}. \quad (13)$$

The number of distinct functions is given by

$$C_X = \prod_{i=1}^{|A|} \binom{|X| - c_X(i)}{h_X(i)}, \quad (14a)$$

$$c_X(i) = \begin{cases} 0, & i \leq 1 \\ \sum_{j=1}^i h_X(j), & \text{otherwise.} \end{cases} \quad (14b)$$

$C_X$  is a multinomial coefficient and can be simplified to

$$C_X = \binom{|X| - c(1)}{h(1)} \binom{|X| - c(2)}{h(2)} \binom{|X| - c(3)}{h(3)} \dots, \quad (15a)$$

$$= \binom{|X|}{h(1)} \binom{|X| - h(1)}{h(2)} \binom{|X| - h(1) - h(2)}{h(3)} \dots, \quad (15b)$$

$$= \frac{|X|!}{h(1)!(|X| - h(1))!} \frac{(|X| - h(1))!}{h(2)!(|X| - h(1) - h(2))!} \frac{(|X| - h(1) - h(2))!}{h(3)!(|X| - h(1) - h(2) - h(3))!} \dots, \quad (15c)$$

$$= \frac{|X|!}{h(1)!h(2)!h(3)! \dots}, \quad (15d)$$

$$= \frac{|X|!}{\prod_{i=1}^{|A|} h(i)!}, \quad (15e)$$

where we for simplicity have neglected to write the subscript  $X$  and in the last term used that  $(|X| - h(1) - h(2) - \dots - h(|A|))! = 1$ . Like the simplified notation for  $h$ , we will also write  $C_i$  for  $C_{X_i}$ .

For the disjoint sets  $\forall_{i \neq j} X_i \cap X_j = \emptyset$ , the functions on  $X_i$  are independent on those on  $X_j, j \neq i$ , and may be chosen independently. Thus, number of functions sharing  $H$  is

$$C_X^{\text{disjoint}} = \prod_i C_{X_i}, \quad (16)$$

where  $C_{X_i}$  is Equation (14) applied to  $h_i$ .

**Example 3.1.** As an example, consider the (ordered) alphabet  $A = \{1, 2, 3\}$  and the histogram  $h_X = [1; 1; 2]$ . Then by Equation (2) we know that  $|X| = 4$ . Finally using Equation (15) we find that

$$C_X = \frac{4!}{1!1!2!} = 12. \quad (17)$$

Assuming that  $X$  is a line, we can list all possible functions which has histogram  $h_X$  as,

$$[0; 1; 2; 2], [0; 2; 1; 2], [0; 2; 2; 1], [1; 0; 2; 2], [2; 0; 1; 2], [2; 0; 2; 1], [1; 2; 0; 2], [2; 1; 0; 2], [2; 2; 0; 1], [1; 2; 2; 0], [2; 1; 2; 0], [2; 2; 1; 0].$$

**Example 3.2.** Another example, for the same alphabet as in Example 3.1 but with  $h_X = [2; 0; 2]$  we follow the same procedure as in Example 3.1 to calculate  $C = \frac{4!}{0!2!2!} = 6$ , and the list possible functions on a linear domain  $X$  as,

$$[0; 0; 2; 2], [0; 2; 0; 2], [0; 2; 2; 0], [2; 0; 0; 2], [2; 0; 2; 0], [2; 2; 0; 0].$$

**Example 3.3.** Continuing Example 2.2 with  $A = \{1, 2, 3\}$ ,  $X = \{1, 2, \dots, 6\}$ ,  $X_1 = \{1, 2, 3, 4\}$ ,  $X_2 = \{3, 4, 5, 6\}$ , and  $h_{\{1\}} = [1; 1; 0]$ ,  $h_{\{1,2\}} = [1; 0; 1]$ ,  $h_{\{2\}} = [0; 2; 0]$ , the number of functions is computed from its non-overlapping parts are

$$C_{\{1\}} = \frac{2!}{1!1!0!} = 2, \quad C_{\{1,2\}} = \frac{2!}{1!0!1!} = 2, \\ C_{\{2\}} = \frac{2!}{0!2!0!} = 1, \quad (18)$$

Thus, the total number of functions for these specific histograms  $h_0$  and  $h_1$  is  $C_{\{1\}}C_{\{1,2\}}C_{\{2\}} = 4$ , and the functions are any combination of

$$f(X'_{\{1\}}) \in \{[1; 2], [2; 1]\}, \quad f(X'_{\{1,2\}}) \in \{[1; 3], [3; 1]\}, \\ f(X'_{\{2\}}) = [2; 2]. \quad (19)$$

One of the 4 functions, which have histograms  $h_1$  and  $h_2$  specified in Equation (12) is thus  $f(X) = f(X'_{\{1\}} \cup X'_{\{1,2\}} \cup X'_{\{2\}}) = [1; 2; 3; 1; 2; 2]$ .

**Example 3.4.** As a final example, consider a one-dimensional function over the alphabet  $A = \{1, 2, 3\}$  and where  $X = X_1 \cup X_2 \cup X_3$ ,  $X_1 = \{1, 2, 3, 4\}$ ,  $X_2 = \{2, 3, 4, 5\}$ ,  $X_3 = \{3, 4, 5, 6\}$ . The unique partitions are then given as,

$$X'_{\{1\}} = \{1\}, \quad X'_{\{1,2\}} = \{2\}, \quad X'_{\{1,3\}} = \emptyset, \quad X'_{\{1,2,3\}} = \{3, 4\}, \\ X'_{\{2\}} = \emptyset, \quad X'_{\{2,3\}} = \{5\}, \quad X'_{\{3\}} = \{6\}. \quad (20)$$

The possible histograms of the singleton domains are

$$h_I \in \{[1; 0; 0], [0; 1; 0], [0; 0; 1]\}, \quad I \in \{\{1\}, \{1, 2\}, \{2, 3\}, \{3\}\}, \quad (21)$$

and for  $X'_{\{1,2,3\}}$ ,

$$h_{\{1,2,3\}} \in \{[2; 0; 0], [1; 1; 0], [1; 0; 1], [0; 2; 0], [0; 1; 1], [0; 0; 2]\}, \quad (22)$$

since  $|X'_{\{1,2,3\}}| = 2$ . The total number of different histograms is,

$$|\mathcal{H}| = \binom{3}{2}^4 \binom{4}{2} = 486. \quad (23)$$

To generate a set of functions and overlapping histograms, we choose a specific set of  $h_I$ ,

$$h_{\{1\}} = [0; 0; 1], \quad h_{\{1,2\}} = [0; 1; 0], \quad h_{\{1,2,3\}} = [1; 1; 0], \\ h_{\{2,3\}} = [0; 0; 1], \quad h_{\{3\}} = [1; 0; 0], \quad (24)$$

and thus,  $h_1 = h_{\{1\}} + h_{\{1,2\}} + h_{\{1,2,3\}} = [1; 2; 1]$ ,  $h_2 = h_{\{1,2\}} + h_{\{1,2,3\}} + h_{\{2,3\}} = [1; 2; 1]$ , and  $h_3 = h_{\{1,2,3\}} + h_{\{2,3\}} + h_{\{3\}} = [2; 1; 1]$ . The number of functions is computed from its non-overlapping parts,

$$C_{\{1\}} = 1, \quad C_{\{1,2\}} = 1, \quad C_{\{1,2,3\}} = 2, \quad C_{\{2,3\}} = 1, \quad C_{\{3\}} = 1, \quad (25)$$

Thus, the total number of functions for these specific histograms  $\mathcal{H} = \{h_1, h_2, h_3\}$  is  $C_{\{1\}}C_{\{1,2\}}C_{\{1,2,3\}}C_{\{2,3\}}C_{\{3\}} = 2$ , and the functions are any combination of

$$f(X'_{\{0\}}) = [2], \quad f(X'_{\{0,1\}}) = [1], \quad f(X'_{\{0,1,2\}}) \in \{[0; 1], [1; 0]\}, \\ f(X'_{\{1,2\}}) = [2], \quad f(X'_{\{2\}}) = [0], \quad (26)$$

and one of the two possible functions sharing  $\mathcal{H}$  is thus  $f(X) = [2; 1; 0; 1; 2; 0]$ .

In the above, we have given a method for generating histograms and functions by repartitioning the domain into disjoint domains. In the following, we will investigate how to find the set of functions, which share a set of overlapping histograms.

## 4. Unique functions from overlapping histograms

For a set of overlapping histograms,  $\mathcal{H} = \{h_1, h_2, \dots, h_k\}$  we have yet to find a closed form solution for counting the number of functions, which share  $\mathcal{H}$ . However, by repartitioning their domain using Equation (7) giving  $|P_k|$  disjoint domains, we are able to recursively calculate the sets of histograms for the repartitioned domains which agree with  $\mathcal{H}$ . For each domain, we have  $\kappa_{h_I}(1, |X_I|)$ ,  $I \in P_k$  different histograms where  $\kappa$  is given recursively as,

$$\kappa_h(j, k) = \begin{cases} 1, & \text{if } k = 0, \\ \sum_{i=l(j,k)}^{u(j,k)} \kappa_h(j+1, k-i), & \text{otherwise,} \end{cases} \quad (27a)$$

$$u(j, k) = \min(h(j), k), \quad (27b)$$

$$l(j, k) = \max\left(0, k - \sum_{i=j+1}^{|A|} h(i)x\right). \quad (27c)$$

**Example 4.1.** For example, given two overlapping subdomains  $X_1$  and  $X_2$ , we repartitioning the domain using Equation (7) into  $X'_{\{1\}} = X_1 \setminus X_2$ ,  $X'_{\{1,2\}} = X_1 \cap X_2$ , and  $X'_{\{2\}} = X_2 \setminus X_1$ .

Further, if  $A = \{1, 2\}$ ,  $X'_{\{1\}} = \{1, 2\}$ , and  $X'_{\{1,2\}} = \{3, 4\}$ , then there are the following possible combinations of histograms for  $h_1$  and  $h_{\{1\}}$ :

$$h_1 = [4; 0] \Rightarrow h_{\{1\}} = [2; 0], \quad (28a)$$

$$h_1 = [3; 1] \Rightarrow h_{\{1\}} \in \{[2; 0], [1; 1]\}, \quad (28b)$$

$$h_1 = [2; 2] \Rightarrow h_{\{1\}} \in \{[2; 0], [1; 1], [0; 2]\}, \quad (28c)$$

$$h_1 = [1; 3] \Rightarrow h_{\{1\}} \in \{[1; 1], [0; 2]\}, \quad (28d)$$

$$h_1 = [0; 4] \Rightarrow h_{\{1\}} = [0; 2], \quad (28e)$$

The recursive evaluation of  $\kappa$  in Equation (27) for this example is visualized as the trees in Figure 2. Not that given  $h_{\{1\}}$ , then  $h_{\{1,2\}}$  is determined directly by Equation (5) as  $h_1 - h_{\{1\}}$ . For example, if  $h_1 = [3; 1] \wedge h_{\{1\}} = [1; 1]$  then  $h_{\{1,2\}} = [3; 1] - [1; 1] = [2; 0]$ .

Given  $\mathcal{H}$ , we can use Equation (27) to sequentially generate a tree of histograms  $h_I$  which agree with  $\mathcal{H}$ . For example, starting with  $h_1$  we can calculate the set of possible histograms for  $(h_{\{1\}}, h_1 \setminus h_{\{1\}})$  pairs. Then for each  $h_1 \setminus h_{\{1\}}$  we calculate the set

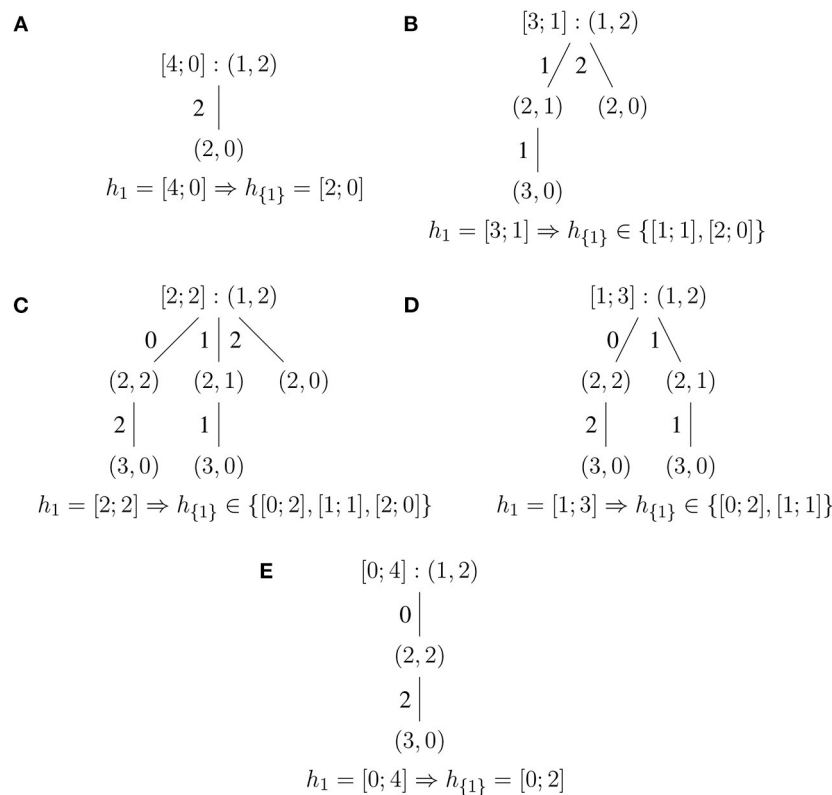


FIGURE 2

Recursive evaluation of Equation (27). (A–E) corresponds to Equations (28a–28e). The nodes are the  $(j, k)$  pair,  $j$  is the index of the following histogram value, and the branch number its value.  $k$  is the number of values still to be decided. The count of leave-values gives the value of  $\kappa$ .



of possible histograms for  $(h_{\{1,2\}}, h_{\{1\}} \setminus h_{\{1,2\}})$  pairs and so on. In practice, we have chosen to implement a sifting algorithm instead, which will be described in the following.

Given a set overlapping domains  $\{X_0, X_1, \dots\}$  and their corresponding histograms  $\{h_{X_0}, h_{X_1}, \dots\}$ , we propose a sifting algorithm that considers a list of candidate functions that are iteratively updated as we consider additional local histograms. We produce candidate functions, and for a particular candidate function  $f$ , which has candidate values at positions  $X^n = \bigcup_{i=0}^{n-1} X_i$ , the next window  $X_n$  and its target histogram  $h_{X_n}$ , we identify yet to be considered region  $X_n \setminus X^n$  and calculate the function

$$g_{X_n \setminus X^n} = h_{X_n} - h_{X_n \cap X^n}. \quad (29)$$

If  $g \geq 0$  and  $\sum_a g(a) = |X_n \setminus X^n|$ , then we cannot refute the candidate, and  $g$  is the histogram of  $X_n \setminus X^n$  which agrees with the histograms  $h_0, h_1, \dots, h_n$ , and hence the candidate  $f$  is replaced with a new set of candidate functions extending  $f(X^n)$  with function values that have histogram  $g$  at  $X_n \setminus X^n$ .

The computational complexity of the algorithm is governed by the sizes of the function, the sizes of  $X_i$ , and the sweeping order of the update of the candidates. In Figure 3 are two unavoidable cases shown for a 2-dimensional domain. In the figure,  $X^n$  are denoted by the blue areas, and  $X_n$  by the green square.

Since each candidate appears to grown binomially by the size that  $X_n \setminus X^n$ , our experiments indicate that a sweeping order, where the cases where  $X_n \setminus X^n$  is small seems to produce fewer maximum number of candidates during the reconstruction. An upper bound on the search tree is given in Section 5. The main part of our algorithm is shown in Figure 4. The full algorithm can be downloaded from github.

**Example 4.2.** An example of a reconstruction is shown in Figure 5, where  $A = \{0, 1, 2\}$ ,  $X = \{0, 1, \dots, 9\}^2$ , and  $X_i$  are  $3 \times 3$  square windows translated in both directions with a stride of 1. In this case, there are two images, which has the same set of local histograms for  $m = 3$ .

## 5. Bound on the size of the search tree

As a measure of the Computational complexity of our sifting algorithm, we will here give an upper bound on the search tree.

Given an  $n \times n$  image with intensities from an alphabet  $A$  and its local histograms  $h_{ij}$  over  $m \times m$  domains,  $X_{ij}$ , where  $m \leq n$ , and where  $ij$  is the lower left corner of the domain. We consider the maximum case of all local  $(n - m + 1)^2$  histograms produced by  $m \times m$  windows translated by 1 over the image domain. Our algorithm considers the histograms in a diagonal order,

$$[h_{11}, h_{21}, h_{12}, h_{31}, h_{22}, h_{13}, h_{41}, \dots, h_{(n-m+1)(n-m+1)}].$$

**Case  $h_{11}$ :** Our sifting algorithm will first produce the set of candidates for  $X_{11}$  which by (15) produces  $\frac{m^2!}{\prod_{i=1}^{|A|} h_{11}(i)!}$  candidates. The pseudo-uniform histogram,  $|h(j) - h(k)| \leq 1$ ,  $j \neq k$  maximizes this value. To prove this, consider two values in the histogram, where  $h(j) > h(k)$ , and the denominator,

$$d = \prod_i h(i)! \quad (30)$$

$$= h(j)!h(k)! \prod_{i \notin \{j,k\}} h(i)!, \quad (31)$$

For a similar histogram  $h'$  which is equal to  $h$ , except  $h'(j) = h(j) - 1$  and  $h'(k) = h(k) + 1$ , then the ratio of their

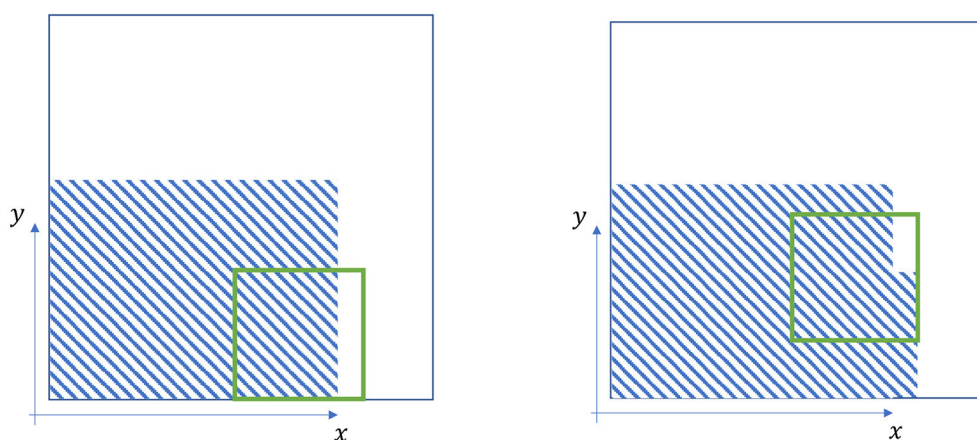


FIGURE 3  
Examples of overlap between a solution candidate (shaded), a new window (green).

```

// H - ('a option * int) [,]
//      a 2d array of all local histograms of mxm patches of
//      an nxn image. The histograms are pairs of values of
//      any type and counts

// Order matters for efficiency, here anti-diagonal ordering
let order =
    H
    |> Array2D.toList // flatten the 2d array to a list
    |> List.sortBy (fun (a,b) -> a+b)
// For all local histograms by order
List.iter (fun (i,j) ->
    let h = H.[i,j]
    // For all the candidate images in lst, iteratively collect updated
    // candidates
    lst <- List.collect (fun (org: 'a option [,]) ->
        let expansion =
            h
            // Make a list of lists of possible updates as
            // coordinate-value tuples
            |> suggest org.[i..i+m0-1,j..j+m1-1]
            // Created a list of images with update
            |> expandSuggestions org i j
        expansion
    ) lst
) order

```

FIGURE 4  
An F# sifting algorithm.

corresponding denominator is,

$$\frac{d}{d'} = \frac{\prod_i h(i)!}{\prod_i h'(i)!} \quad (32)$$

$$= \frac{h(j)!h(k)!}{(h(j)-1)!(h(k)+1)!} \quad (33)$$

$$= \frac{h(j)}{h(k)+1} \geq 1. \quad (34)$$

When  $h(j) = h(k) + 1$  then  $\frac{d}{d'} = 1$  otherwise  $\frac{d}{d'} > 1$ , and we conclude that  $d$  is minimized for pseudo-uniform histograms. Since  $m^2$  is a constant,  $\frac{m^2!}{\prod_{i=1}^{|A|} h(i)!}$  is maximized for pseudo-uniform histograms. Writing  $m^2$  by its integer quotient and remainder,

$$m^2 = q|A| + r \quad (35)$$

where  $q$  and  $r$  are whole numbers and  $0 \leq r < |A|$ , then the pseudo-uniform histogram will have  $|A| - r$  bins with  $q$  values and  $r$  bins with  $q + 1$ , and the largest number of candidates for the left-most part of the image is

$$\frac{m^2!}{\prod_{i=1}^{|A|} h_{11}(i)!} = \frac{m^2!}{q!^{|A|-r}(q+1)!^r}. \quad (36)$$

**Case  $h_{21}$ :** Our algorithm next considers the histogram  $h_{21}$  for the window  $X_{21}$ , which is a translated 1 wrt.  $X_{11}$ , i.e.,  $|X_{11} \setminus X_{21}| = |X_{21} \setminus X_{11}| = m$ . When  $h_{X_{11} \setminus X_{21}} = h_{X_{21} \setminus X_{11}}$ , then none of the candidates generated by  $h_{11}$  can be discarded, and for each, we must consider all the additional candidates for  $h_{X_{21} \setminus X_{11}}$ .

In the worst case,  $h_{X_{21} \setminus X_{11}}$  is pseudo-uniform. Writing  $m$  in terms of its integer quotient and remainder,

$$m = p|A| + s \quad (37)$$

where  $p$  and  $s$  are whole numbers  $0 \leq s < |A|$ , this gives us

$$\frac{m!}{\prod_{i=1}^{|A|} h_{X_{21} \setminus X_{11}}(i)!} = \frac{m!}{p!^{|A|-s}(p+1)!^s}. \quad (38)$$

additional hypotheses to consider for each existing candidate.

**Case  $h_{12}$ :** Having reach this histogram, all candidates agree with  $h_{11}$  and  $h_{21}$ . Since,  $|X_{12} \setminus (X_{11} \cup X_{21})| = |X_{12} \setminus X_{11}| = m$ , the number of additional hypotheses for each candidate are the same as derived for case  $h_{21}$ .

**Case  $h_{22}$ :** Having reach this histogram, all candidates agree with  $h_{11}, h_{21}, h_{12}, h_{31}$ . Since,  $|X_{22} \setminus (X_{11} \cup X_{21} \cup X_{12} \cup X_{31})| = |X_{22} \setminus (X_{11} \cup X_{21} \cup X_{12})| = 1$ , and there is at most one solution for this solution corresponding to a non-negative value in difference between  $h_{22} - h_{X_{22} \cap (X_{11} \cup X_{21} \cup X_{12})}$ . If this histogram difference is has negative value, then the candidate solution can be discarded, however, for simplicity's sake, we will ignore this. Thus, this case does not give additional candidates.

**Bound on the number of candidate images:** By the anti-diagonal order, we 1 time are in Case  $h_{00}$ ,  $n - m$  times in Case  $h_{i,1}, i > 1$  and in Case  $h_{1,j}, j > 1$ , which are identical to Cases  $h_{2,1}$  and  $h_{1,2}$ , and  $(n - m - 1)^2$  times in Case  $h_{i,j}, i, j > 1$ , which are identical to Case  $h_{22}$ . Thus, we conclude that the

```

Original image:
[[0; 0; 0; 2; 0; 0; 0; 2; 1; 2]
 [0; 2; 0; 0; 0; 0; 1; 0; 1; 0]
 [0; 2; 1; 0; 2; 2; 0; 1; 0; 2]
 [0; 0; 2; 2; 0; 0; 2; 0; 2; 0]
 [2; 2; 0; 0; 2; 1; 1; 0; 1; 0]
 [1; 2; 1; 2; 0; 0; 1; 2; 2; 0]
 [2; 1; 2; 1; 1; 2; 2; 1; 2; 2]
 [1; 0; 1; 1; 0; 0; 2; 1; 0; 1]
 [0; 0; 0; 2; 0; 0; 2; 2; 1; 0]
 [1; 2; 1; 1; 0; 0; 2; 1; 1; 0]]

metameric images:
[[[0; 0; 0; 2; 0; 0; 0; 2; 1; 2]
  [0; 2; 0; 0; 0; 0; 1; 0; 1; 0]
  [0; 2; 1; 0; 2; 2; 0; 1; 0; 2]
  [0; 0; 2; 2; 0; 0; 2; 0; 2; 0]
  [2; 2; 0; 0; 2; 1; 1; 0; 1; 0]
  [1; 2; 1; 2; 0; 0; 1; 2; 2; 0]
  [2; 2; 1; 1; 2; 1; 2; 2; 1; 2]
  [1; 0; 1; 1; 0; 0; 2; 1; 0; 1]
  [0; 0; 0; 2; 0; 0; 2; 2; 1; 0]
  [1; 2; 1; 1; 0; 0; 2; 1; 1; 0]];
 [[0; 0; 0; 2; 0; 0; 0; 2; 1; 2]
  [0; 2; 0; 0; 0; 0; 1; 0; 1; 0]
  [0; 2; 1; 0; 2; 2; 0; 1; 0; 2]
  [0; 0; 2; 2; 0; 0; 2; 0; 2; 0]
  [2; 2; 0; 0; 2; 1; 1; 0; 1; 0]
  [1; 2; 1; 2; 0; 0; 1; 2; 2; 0]
  [2; 1; 2; 1; 1; 2; 2; 1; 2; 2]
  [1; 0; 1; 1; 0; 0; 2; 1; 0; 1]
  [0; 0; 0; 2; 0; 0; 2; 2; 1; 0]
  [1; 2; 1; 1; 0; 0; 2; 1; 1; 0]]]

Candidate images per iteration:
252 846 2820 3288 2274 7358 5103 4554 3393 14616 58320 28647 16956 5073
6003 22923 10458 4734 1530 516 804 2412 1068 1050 120 120 120 72 216
144 144 72 72 72 72 216 144 144 72 72 48 48 48 48 24 24 12 12 4 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2

```

FIGURE 5

Reconstructing an image from its local histogram with  $m = 3$ . In this example, the solution set contains two images.

worst case scenario is reached when all histograms considered are pseudo-linear, in which case a maximum of

$$\frac{m^2!}{q!|A|^{-r}(q+1)!^r} \left( \frac{m!}{p!|A|^{-s}(p+1)!^s} \right)^{2(n-m)}$$

hypotheses must be considered. Examples of the number of hypotheses by the above equation for a small set of  $n$ ,  $A$ , and  $m$  are given below

$n$	$ A $	$m$	#Hypotheses
10	2	2	393216
10	2	3	602654094
10	3	2	786432
10	3	3	131651795681280
11	2	2	1572864
11	2	3	5423886846
11	3	2	3145728
11	3	3	4739464644526080

**Note:** In practice, the number of hypotheses in memory is considerably smaller. Consider the case of  $n = 3$  and  $m = 2$ .

The initial 3 histograms  $h_{11}$ ,  $h_{21}$ , and  $h_{12}$  generates hypotheses for the  $m^2 - 1 = 3$  values in  $|X_{22} \cap (X_{11} \cup X_{21} \cup X_{12})|$ , for which there are only different  $\binom{3+|A|-1}{|A|-1}$  histograms, and  $h_{22}$  must be  $A^{\binom{3+|A|-1}{|A|-1}}$  out of the possible  $\binom{4+|A|-1}{|A|-1}$  histograms. Similarly, if the histogram difference contains zero-values, then any candidate, which has a non-zero value at the corresponding histogram point can be discarded. This happens often, when  $m^2 < |A|$ .

## 6. Conclusion

In this article, we have considered locally overlapping histograms of functions from discrete domains and codomains of any dimension. Histograms of signals and images have been studied in the literature extensively, and particularly, the seminal work on locally orderless images (Koenderink and Doorn, 1999), the notion of local histograms has gained a solid theoretical basis. The authors' work map discrete functions and histograms into the continuous domain, in a manner that

makes differentiation of discrete functions well-posed. Their work, however, left the essential question unanswered: what is the expression power of histograms? In [Sporring and Darkner \(2022\)](#) a partial answer is given to this question for binary signals with densely overlapping histograms, and in this article, we extend this work for non-binary discrete functions of any dimension and with windows of any overlap, and in this article we have:

- Presented a simple algebra for histograms of discrete domain and co-domains based on non-overlapping sets.
- For a given set of covering sets in the domain, we have given a constructive method for identifying unique, non-overlapping sets which cover the domain.
- We have given an equation for the size of the set of all possible histograms based on the set of unique, non-overlapping domain sets.
- For a specific set of histograms of the individual unique, overlapping and non-overlapping sets, we have given
  - an equation for calculating the corresponding histograms of any set in the domain and
  - an equation for counting the total number of functions with these histograms.
- Presented an algorithm, which given a set of overlapping histograms, produce the set of functions, which share these histograms.

Understanding the expression power of local histograms is not done. For one, we still seek to connect the results obtained for discrete functions with the continuous domain.

## References

- Gabor, D. (1946). Theory of communication. Part 1: the analysis of information. *J. Inst. Electr. Eng.* 93, 429–441. doi: 10.1049/ji-3-2.1946.0074
- Koenderink, J. J., and Doorn, A. J. V. (1999). The structure of locally orderless images. *Int. J. Comput. Vis.* 31, 159–168.
- Sporring, J., and Darkner, S. (2022). Reconstructing binary signals from local histograms. *Entropy* 24, 433. doi: 10.3390/e24030433
- Stover, C. (2022). *Finite Additivity*. From MathWorld—A Wolfram Web Resource, created by Eric W. Weisstein. Available online at: <https://mathworld.wolfram.com/FiniteAdditivity.html>
- Wu, Y. N., Zhu, S. C., and Liu, X. (2000). Equivalence of julesz ensembles and frame models. *Int. J. Comput. Vis.* 38, 247–265. doi: 10.1023/A:1008199424771

## Data availability statement

The code is publicly available at: <https://github.com/sporring/reconstructionFromHistograms>.

## Author contributions

JS and SD: conceptualization and writing—review and editing. JS: formal analysis, methodology, software, and writing—original draft. Both authors have read and agreed to the published version of the manuscript.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher. All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.





## OPEN ACCESS

## EDITED BY

Radhya Sahal,  
Independent Researcher, Cork, Ireland

## REVIEWED BY

Hager Saleh,  
INSERM U1166 Unité de Recherche  
sur les Maladies Cardiovasculaires, du  
Métabolisme et de la Nutrition, France  
Faiza Alssaedi,  
University of Galway, Ireland

## \*CORRESPONDENCE

Ella Peltonen  
✉ ella.peltonen@oulu.fi

## SPECIALTY SECTION

This article was submitted to  
Mobile and Ubiquitous Computing,  
a section of the journal  
Frontiers in Computer Science

RECEIVED 31 July 2022

ACCEPTED 05 December 2022

PUBLISHED 21 December 2022

## CITATION

Peltonen E (2022) The role of gender  
in the International Conference on  
Pervasive Computing and  
Communications.  
*Front. Comput. Sci.* 4:1008552.  
doi: 10.3389/fcomp.2022.1008552

## COPYRIGHT

© 2022 Peltonen. This is an  
open-access article distributed under  
the terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other  
forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which  
does not comply with these terms.

# The role of gender in the International Conference on Pervasive Computing and Communications

Ella Peltonen\*

Empirical Software Engineering in Software, Systems and Services (M3S), University of Oulu, Oulu, Finland

The International Conference on Pervasive Computing and Communications (IEEE PerCom) is a CORE 2021 A\* conference (top 7% of ranked venues) that aims to present scientific advances in a broad spectrum of technologies and topics in ubiquitous/pervasive computing, including wireless networking, mobile and distributed computing, sensor systems, ambient intelligence, and smart devices. During the last couple of years, the PerCom organization committee has successfully included many prestigious female researchers to submit, participate, and organize the conference. However, there is still work to do and to help the progress, this article analyses the history of the conference from a gender perspective. This article goes through accepted articles of the last 20 years of the PerCom conferences, showing that even if the role of female authors, in general, has increased, more first and leading female researchers should still be welcomed in the community. Through this analysis, this article aims to highlight the role of gender in the conference program and seeks to find trends and possible improvements to achieve a broader gender balance in pervasive computing.

## KEYWORDS

pervasive computing, ubiquitous computing, mobile computing, gender, women

## 1. Introduction

It is estimated that the partition of the female researchers in computer science is approximately between 15 and 30% (Frachtenberg and Kaner, 2022), which is in one of the worst ratios even among the other STEM fields (Holman et al., 2018; Wang et al., 2021). Some recent bibliographic studies end up even slower numbers, giving only 10% on average for female authorship (Mattauch et al., 2020; Frachtenberg and Kaner, 2021). However, most studies suggest that there is an increasing trend in female participation and authorship over years—even if it is growing slowly. For example, among the prestigious ACM conferences between 1967 and 2007, 10–44% of authors were female with an increasing trend (Cohoon et al., 2011). However, variation between the sub-fields in computer science has been reported to be still considerable. Female authors have been seen as more active in the stereotypical “soft” side of the field, including human-computer interaction, management, design, and other areas involving

human factors. Men, on the other hand, dominate stereotypical “hard” sub-fields of algorithms, reliability, and performance (Cohoon et al., 2011).

Interestingly, pervasive and ubiquitous computing are fields that combine certain sides of both “soft” and “hard” computer sciences. Wherever there are traditionally appreciated large-scale field studies involving real users, there is also a strong consensus on performing measurable data analytics, applying the latest algorithmic inventions, and bringing together various sub-fields inside computer science. There are two key conferences (among various other venues, journals, and workshops) that can be immediately named in the field of pervasive and ubiquitous computing: the International Conference on Pervasive Computing and Communications (IEEE PerCom) and the ACM International Joint Conference on Pervasive and Ubiquitous Computing (ACM UbiComp). Together, they represent also the two key publisher-community organizations in computer science. Both have also immersive history, UbiComp running from 1999 and PerCom since 2003. Both conferences have a strong commitment to their respective and host organizations’ equality, diversity, and code of conduct statements. They are also well-established in their practices for aiming to increase female participation, including inviting diverse technical program and organization committees, involving prestigious female keynote speakers, and hosting the N2Women (Networking Networking Women<sup>1</sup>) networking event yearly.

Considering this history, it is indeed of substantial importance to find numerical evidence for the success of these efforts to develop female participation in pervasive and ubiquitous computing. For some reason, ACM UbiComp was not considered in a study by Cohoon et al. (2011) that highlighted female participation in some other major ACM conferences 10 years ago. In a more recent study in Bonifati et al. (2022) focus on female presentation in database community. They are able to conclude with very similar results to Cohoon et al. 10 years ago: the presentation of female authors is slowly increasing, but women are still an underrepresented group in computer science. Similarly, there is variation between conferences, again female researchers are more prominent in human factor-related areas than strictly technical fields.

This article focuses on finding the trends in female authorship in pervasive and ubiquitous computing. For simplicity of the research work, the focus is given only to the IEEE PerCom community, where the author is an active member. However, in the future, a comparison between these two conferences would indeed provide additional insights. The research questions and/or hypotheses this article focuses on are:

1. If and how much female researchers are underrepresented in the IEEE PerCom community?
2. Can we see any prominent trends in female authorship inside the community when analyzing female first authors and leading authors, or female authors in general?
3. Can we identify if there are an underlying stereotypical distribution of research topics inside the community, as reported in the previous work from other computer science fields? i.e., female researchers focusing on more human factors than corresponding male researchers?

## 2. Materials and methods

### 2.1. Dataset

The data consists of all accepted articles of the International Conference on Pervasive Computing and Communications (IEEE PerCom), from its beginning in 2003 to 2022 when the conference celebrated its twentieth anniversary. For limiting the dataset, only accepted main track (full and concise) articles were considered. The conference also includes yearly various workshops, demonstrations, Ph.D. forum posters, and work-in-progress papers, which were not included for the sake of the number of articles. The limitation was also motivated by the fact that the PerCom main track is the most competed part of the conference, where only 15–20% submitted articles are accepted yearly.

The dataset includes publicly available information about the accepted articles, including the title of the article, the full names of the authors, and the year of the conference where the work was presented. The data sources used to collect this information were the IEEE Explore database<sup>2</sup> and the PerCom websites<sup>3</sup> from the years accepted article information is available. All the PerCom conferences have their proceedings in the IEEE Explore database. Requesting the title and author information does not require an IEEE subscription or membership. The data were fetched from these sources by hand. In total, the data entries consist of 610 items.

### 2.2. Delivering the gender

In this article, the gender of each author is delivered by following the next procedure. If the name is a traditional (usually western) female name, the author is considered female. If the name is such that can be used for any gender, does not associate with a specific gender, or is rare enough to be unsure, the person behind the name is found out *via* Google Search. In most cases, researchers have a website, either personal or

<sup>1</sup> <https://n2women.comsoc.org/>

<sup>2</sup> <https://ieeexplore.ieee.org/Xplore/home.jsp>

<sup>3</sup> <https://www.percom.org/>

institutional, with a biography that specifies also the English pronouns they are using. If the pronoun is “she,” the author is considered female. There were no cases met where the author would have preferred a neutral pronoun “they” or equivalent, so the situation of these authors did not need to be solved (if so, compounding non-binary authors with female authors would have to be considered). In rare cases, the pronouns could not be found (for example, the author has no website for some reason or other) whenever the name was considered female or male based on the online name finder<sup>4</sup> and if not found, a general Google search -based consensus.

Any type of method of delivering the gender of the authors does not come without limitations. This work of finding the authors’ genders was all done manually by hand, so it is prone to human error. However, even the automated methods of inferring names include biases (Karimi et al., 2016). Identification of the international names that are not gender-specific is prone to incorrect categorization and pronouns used publicly for professional websites may not reflect the true identity of the gender, being a more complex phenomenon than a simple two-class category.

## 2.3. Authorship definitions

This work considers three categories of authorship: if there is a female author at all, if the first/main author is a female, and if the last/leading author is a female. This diversion is already used in some previous studies (Bonifati et al., 2022). First, overall authorship, i.e., if there is at least one female in the group of authors for the given article. These articles are considered in the *total* number of articles with female authors, subsequently in this work. Please note that even if there is more than one female author, the article is counted only once.

Second, the *first author* is simply defined as the first name of the author list provided. In the PerCom community, there is a tradition that the first author contributed most significantly, holds the correspondence, and usually also presents the work at the conference. Third, the *lead author* is considered the last author of the article. In the PerCom tradition, this place is usually reserved for the group leader, professor, or the person who takes the most senior position in the work and has the highest responsibilities for the presented results. All-female authored articles (single or several authors) were almost non-existing, thus they are not considered as a single group.

## 2.4. Ethical considerations

The data utilized in this work is collected from publicly available databases, such as IEEE Explore and conference

websites, and the author’s personal/institutional websites. There is no information that would not have been publicly available, and for sake of combining such information in this work, no individual authors, their gender, or affiliation, are discussed or revealed. The IEEE PerCom steering committee has been made aware of the ongoing study, but it has not influenced the data collection, methodology, or results. The data management and processing pipelines are subject to the University of Oulu Internal Ethical Board, and follow the ethical guidelines of the Finnish National Board on Research Integrity (TENK).

## 3. Results

### 3.1. Gender in accepted articles

Out of 610 accepted articles during 20 years of PerCom, 203 had at least one female author (33%). This number alone bases it in the high end of the reported 10–44% of computer science conferences in Cohoon et al. (2011). In comparison to newer studies, average in computer systems was reported to be only around 10% in Frachtenberg and Kaner (2022) and 30–70% in database research reported in late 2021 (Bonifati et al., 2022). The distribution of the number of the PerCom articles over the years is illustrated in Figure 1.

PerCom gained its highest female presentation, 56% in 2020, just before the COVID-19 pandemic (the articles for the next year’s conference are submitted around September the previous year, i.e., those accepted for the 2020 conference were ready for evaluation in September 2019). In 2021, the number of accepted articles with female authors was 42% and in 2022 only 38%. Even if the COVID-19 pandemic has reportedly affected negatively, especially to the careers of the female researchers (Inno et al., 2020; Deryugina et al., 2021), there is, unfortunately, no quantitative or qualitative evidence in the data that this was the case with the PerCom conference specifically.

The distribution of the articles with female first and lead authors can also be seen in Figure 1. Please note that the amount of any female author always includes both cases of first or lead female author, thus being higher by definition. Out of 610 articles total, 73 had a female first author (11.9%) and 69 had a female lead author (11.3%). In comparison to the database field where the numbers have been reported similarly (Bonifati et al., 2022), the average for female first authors varies between 12 and 25% and female lead authors between 15 and 25%. It is noteworthy that even if the number of overall female-authored articles is higher in PerCom than in some other conferences (e.g., in systems research) and equal to, for example, the database field, the number of first and lead female authors is hindering behind significantly.

<sup>4</sup> <https://gendernamefinder.com/>

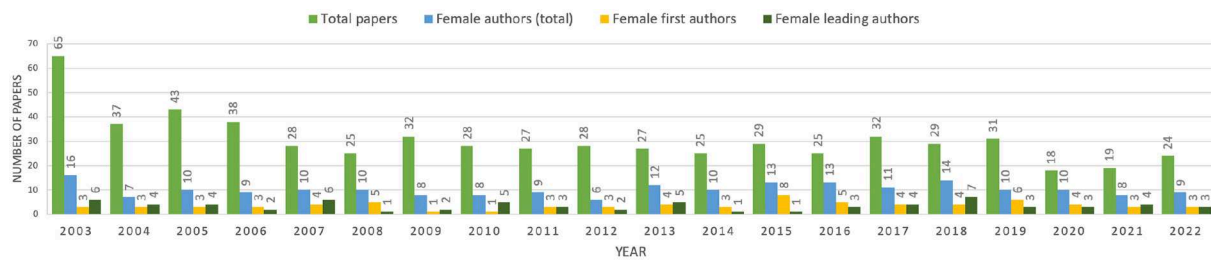


FIGURE 1

Number of accepted articles in PerCom from 2003 to 2022, and distribution (absolute numbers) of female-authored articles, female first authors, and female lead authors.

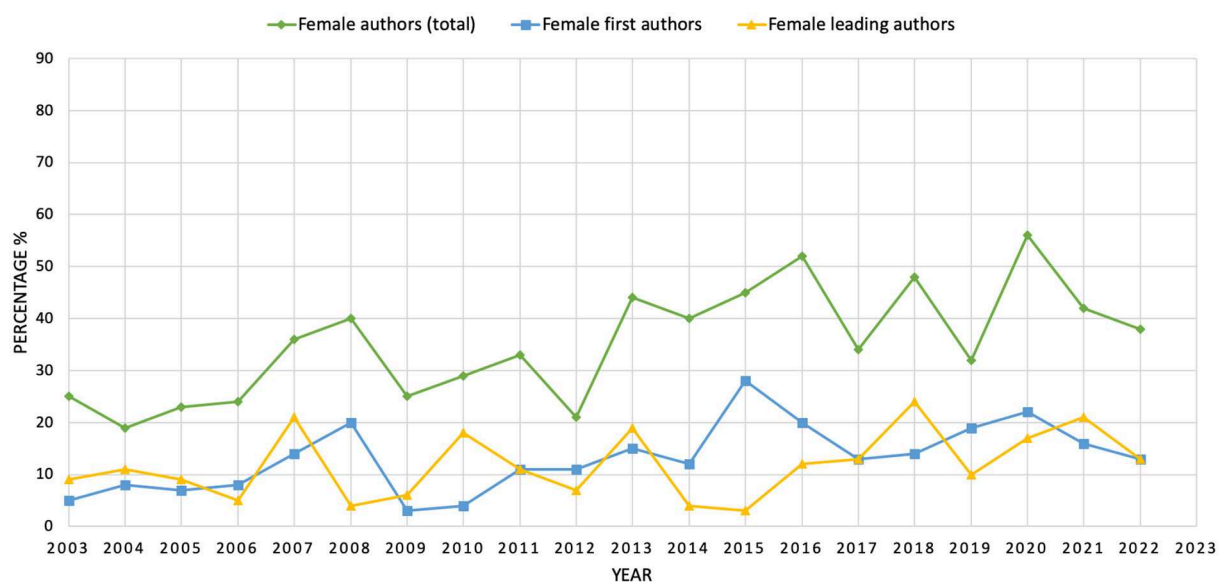


FIGURE 2

Percentages of female authors (total, first, or as lead authors) in PerCom from 2003 to 2022.

### 3.2. Gender trends in authorship

By hypothesis, it is expected that when awareness of gender equality and more equal opportunities arise, also the number of female researchers should increase over time somewhat linearly. Some long-term studies have already supported such a hypothesis (Cohoon et al., 2011) even if the speed of the development seems to be, indeed, slower than one would expect. To analyze the authorship trends in the PerCom conference, the absolute numbers of female authors are converted into percentages of all accepted articles of the corresponding year. Please note that PerCom has no hard limit on how many articles should be accepted yearly, and the number has varied drastically from the beginning of the conference, from approximately 40 articles to the current 20 articles per year.

The percentages of female authors (total, first, or as lead authors) in PerCom from 2003 to 2022 are shown in Figure 2. An overall glance at the trends as well as the removal of the seasonal variation (i.e., removing the polynomial trend from the data series of the total female authors) supports the hypothesis of a possible linear trend. Thus, the data series can be fitted with the linear regression model ( $y \sim 1 + x1$ ). Figure 3 shows the results of the regression analysis, for each category (all female authors, female first authors, and female lead authors), correspondingly.

Figure 3A shows that the female authorship in general is in upward trend—steady, but not fast, with statistical significance ( $R^2 = 0.46, p = 0.00$ ). The trend of female first authors is also increasing but less steadily ( $R^2 = 0.33, p = 0.007$ , see Figure 3B) than the female authors in general. However, the role of leading female authors has not significantly changed from the past 20 years ( $R^2 = 0.125, p = 0.126$ ) as seen in Figure 3C. Even if more



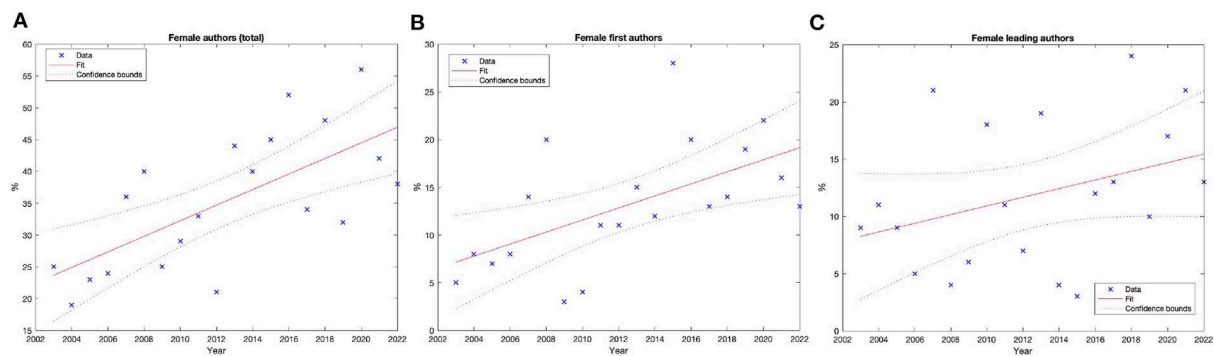


FIGURE 3

Linear regression fits for percentages of (A) female authors ( $R^2 = 0.46, p = 0.00$ ), (B) female first authors ( $R^2 = 0.33, p = 0.007$ ), and (C) female leading authors ( $R^2 = 0.125, p = 0.126$ ) out of total number of accepted articles.

female authors have been attracted to the PerCom community, they take the role of “junior” positions either as first authors or somewhere in the middle of the author list. However, it is noteworthy that there are several “mid-term” female professors in the field, which may lead to a situation where the first name is reserved for a more mature male professor even if the “second last” position included tasks of a leading author. These cases are not analyzed in this article, simply because either the PerCom articles or IEEE database list the roles or responsibilities of the authors.

### 3.3. Gender and research topics

Computer science and other STEM fields are not free of gender stereotypes (Thébaud and Charles, 2018). These include traditional stereotypes like “men are strong and competent” and “women are softer and more human-oriented.” Luckily, there are already arguments that the stereotype of men being more intelligent than women has already started to vanish (Eagly, 2018). Even if different personalities and skill sets are required in computer science, it can be harmful to analytical women and more emotional men to be subject to such stereotypes during hiring processes (Thébaud and Charles, 2018), when choosing a graduate or postgraduate program (Ertl et al., 2017), and other occurrences also in research. The stereotypes have been linked to the shortage of women in the technical and other STEM fields, mainly through students’ and kids’ tendency of choosing other fields over STEM under the biased image of the field (Piatek-Jimenez et al., 2018).

The stereotypes in the research field of computer science can be, on a general level, summarized as the tendency of women to focus on areas involving human-factors whereas male researchers are seen to success in more technical topics (Cohoon et al., 2011). To analyze this in the context of the

PerCom conference, we take a look at the titles of the published articles. The hypothesis that can be considered here is that the titles should be possible to organize into groups through topic modeling. These groups, when later analyzed with authors’ gender information, should show if certain research topics are, indeed, more prominent among the female researchers than their male counterparts.

The topic modeling algorithm chosen is a commonly known Dirichlet allocation (LDA). The pipeline of the LDA analysis is to choose the representative words among the article titles and then associate the words with characterizing topics. As a result, an approximation is given that how important each topic is for a certain title. The preprocessing of the article titles involves a standard procedure of lowering all the cases and then removing any punctuation (including colons and hyphens), words <2 characters, and so-called stop-words (including prepositions and conjunctions). For the LDA hyper-parameter training (i.e., how many topics should be expected), goodness-of-fit is calculated by the perplexity of a held-out set of items i.e., the article titles. The hyper-parameter training finishes (i.e., ends with the lowest perplexity) with approximately 20 topics being an optimal number for this dataset.

The results of the LDA model fit are given in Figure 4, where each topic is visualized as a word cloud. The highlighted words are those of the highest probability in each topic, and the size of the words corresponds to their probability, too. For reference, the PerCom conferences’ predefined topics or themes can be found in the year call for papers<sup>5</sup>. The topics from the LDA model characterize well the PerCom community’s interests: mobile sensing (topic 1), sensor networks (topic 3), systems (topic 4), human-activity recognition (topic 6), wearable computing (topic 7), indoor localization (topic 8),

<sup>5</sup> <https://www.percom.org/call-for-papers/>



authentication and RFIDs (topic 9), *ad hoc* networks (topic 13), and smart homes (topic 18), to highlight some.

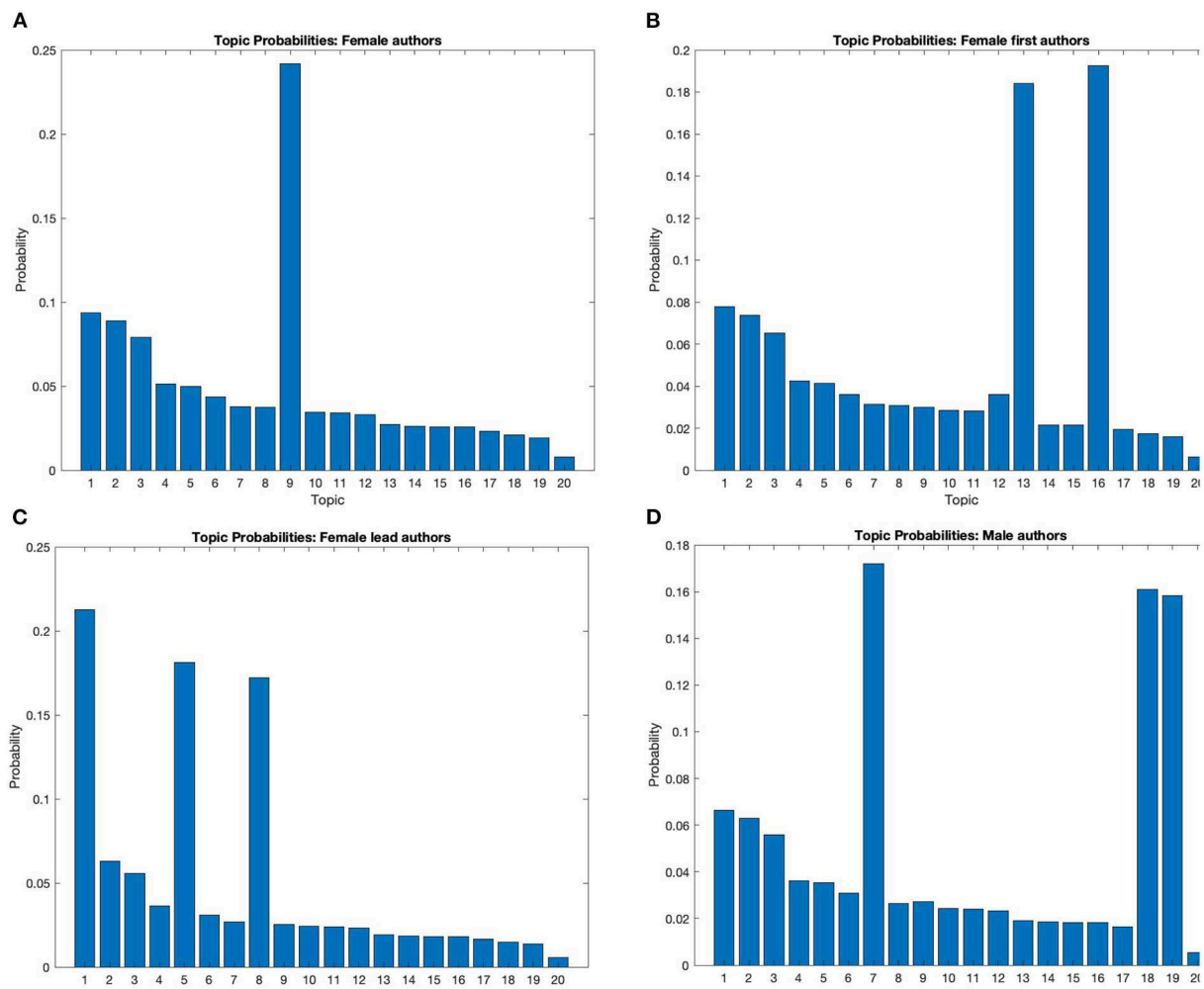
After finding the models, the next step is to associate authors' genders with the topics. This is done by separating the titles into four groups based on their authorship status: a) at least one female author, b) female first author, c) female lead author, and d) only male authors. The results are shown in Figures 5A–D, respectively. The numbers on the x-axis represent the topics given by the LDA model (see Figure 4 for reference) and the y-axis represents the probability of each topic being present in the titles of each group.

Surprisingly enough, the female authors of the PerCom conference seem NOT to follow the stereotype of focusing on research topics driven by human factors (see Figure 5A). The most “female” topic number 9 includes keywords such as “context” and “approach,” followed by words such as “digital,” “continuous,” “multimodal,” “signals,” and “camera.” These topics can be identified as data analytics, sensor data

processing, and in general pervasive and data-driven approaches that are usually highly technical and mathematical. For the female first authors, the most prominent topics are numbers 13 and 16 (see Figure 5B). Topic 13 can be defined by network management, routing, and *ad hoc* networks—all extensively technical keywords. Similarly, topic 16 involves keywords such as “frameworks” and “energy efficient.” No human-factors present.

For female leading authors, the leading topics are numbers 1, 5, and 8 (see Figure 5C). Here highlighted are keywords including themes of mobile sensing, protocols and service discovery, and indoor localization. It is noteworthy that due to the small number of female leading authors in general, the results might be biased toward individual professors' research interests. However, they are members of the community.

To compare, the results of the male-only articles are shown in Figure 5D. Here the highlighted topics are numbers 7, 18, and 19. These topics involve keywords indicating wearable



**FIGURE 5**  
Probability distribution of (A) all female authors, (B) first female authors, (C) lead female authors, and (D) male authors among the 20 topics. Each topic is represented as a number in the horizontal axis; see Figure 4 for the topic descriptions.

computing and smartphones, smart home, and crowd-sourcing. Especially, research on smartphones and smart homes includes user perspectives i.e., human factors, which is unlikely the stereotypical hypothesis would predict. Thus, it is indeed interesting to conclude that the stereotypical perspectives seem not to play a role in the PerCom community research topics or how to research topics are chosen in general within the community. Indeed, as a personal experience from the technical program committee meetings, the articles are discussed through their overall scientific value instead of the topics they address, as long as those topics are within the PerCom conference's interests that include user experiences and human factors. However, it is possible that because there are other prominent and high-class venues in close reach to the PerCom community (including conferences like UbiComp and CHI, and multiple journals), the most user-focused

or qualitative works do not become submitted to the PerCom conference.

## 4. Discussion

### 4.1. About the obstacles

When studying female presentation in computer science, the main question arises what could we do to improve the situation? Various activities have taken place, over the specialized networks such as N2Women and inside the conference committees themselves, especially rising awareness of the female representation in different panels, committees, and reviewer boards, as well as keynote speakers (Martin, 2014). Focusing on the systems research in computer science, Frachtenberg and Kaner (2022) studied the conference factors

and their influence on gender diversity: size of the conference, whether the double-blinded review was applied, diversity in conference organizational roles, and different diversity initiatives. However, they concluded none of these factors influenced significantly the gender ratio. Similarly, Bonifati et al. (2022) did not find an effect on double-blind review over single-blind review even if they had data where the same conference had made the change in their review process. Both works (being from 2022) however concluded, that there is a chance that some of the effects are only visible after the next decade or so, when the current Ph.D. students become leaders of their own labs, for example.

Of course, some obstacles to women's scientific careers are larger than a single conference can fix (Huang et al., 2020). These include, but are not limited to, a higher drop-out rate for various career-related reasons, and fewer resources and changes to building an effective "paper factory" as a senior researcher. Even if being a female may not affect negatively in peer-review process (Tomkins et al., 2017), female professors and senior researchers face a high load of faculty services and teaching (Misra et al., 2012; Roper, 2019) that is immediately away from the productive research time. Huang et al. (2020) conclude that the most pronounced—and also the most worrisome—gender gap is indeed between the most productive authors. Those "leading" researchers should be the role models for the next generations, too (Bettinger and Long, 2005). This is comparable to the results shown on leading authors' increase rate over year: even if there are more female authors and even more female first authors present, the number of leading female authors is barely increasing.

## 4.2. Concrete actions

To summarize, the PerCom conference seems to be on a positive track in attracting female authors as long as they work together with their male peers. Female-only authored articles are still consistently missing, and the role of female researchers in first and leading authorship positions is still scarce. To address this, more effort should be provided in attracting female Ph.D. researchers to submit as first authors and participate in the conference in general to find the community. Possible concrete actions include an international female junior researcher fellowship aimed to fund participation in the conference. This should complement the N2Women Fellowship that funds only a single (usually female) person to organize the actual N2Women meeting within the conference.

The data utilized in this work is limited to the accepted articles because the names and titles of the submitted but rejected articles remain confidential information. Because PerCom follows the double-blinded peer-review process, it is hard to implicitly guide toward more female-friendly peer-review progress or establish concrete actions in this area. However, it is important to address the situation of the female

professors and research leaders and how they see the PerCom community as a potential publication and discussion forum. For this, a possible concrete action would be to run a questionnaire study to gather the information that is now missing in the data: why do female authors submit or do not submit to the conference? How do they see the atmosphere at the conference? How equality is addressed in practice, and is there unknown obstacles that should be identified and addressed?

The PerCom Diversity, Equity, and Inclusion commitment listed on the website is somewhat generic<sup>6</sup>. It states that "PerCom is committed to providing an equitable and inclusive forum that supports these rights for all" but does not list actions on how equality and inclusive participation are guaranteed. The statement continues: "We want every participant to feel welcome at the conference. We aim to provide a safe, respectful, and harassment-free conference environment for everyone." As usual in these statements, there is an anonymous email that can be addressed if "behavior inconsistent with such principles" is met, but by personal experience, few are confident to send such an email without strong concrete evidence of a serious harassment case. Actions should be taken before the community even has to address such situations. Thus, another concrete action could be to address not only the prevention of clear harassment but also how to provide a positive environment and support equality in everyday actions regarding the community's operations, such as peer-review process, conference attendance, and even after parties. The ending of the COVID-19 pandemic and returning to the "new normal" will provide a great opportunity to address also questions of equality and inclusiveness in conference organization.

## 4.3. Conclusion

Female researchers are still represented in authorship of the scientific papers in the STEM field, including the IEEE PerCom conference discussed in this paper. In this paper, we have presented that there is an upward trend of female participation in the published articles of the last 20 years of PerCom. This trend is especially prominent when considering female authors in general or female first authors, i.e., young or early-career researchers. However, the trend is not equally strong with leading authors, i.e., among established professors and research leaders. In addition to the trend analysis, we studied if there are underlying stereotypical distribution of research topics inside the community, as reported in the previous work from other computer science fields. Here, we can conclude that the female researchers in the IEEE PerCom community are not focusing on more "soft" human factors than corresponding male researchers, as suggested by similar studies in other fields. Indeed, they are active in various technical topics found in pervasive computing.

<sup>6</sup> <https://www.percom.org/percom-diversity-equity-and-inclusion-statement/>



## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: IEEE Explore.

## Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

## Acknowledgments

The author is thankful to the PerCom community as a whole for providing such a great venue for scientific discussion.

## References

- Bettinger, E. P., and Long, B. T. (2005). Do faculty serve as role models? The impact of instructor gender on female students. *Am. Econ. Rev.* 95, 152–157. doi: 10.1257/000282805774670149
- Bonifati, A., Mior, M. J., Naumann, F., and Sina Noack, N. (2022). How inclusive are we? *ACM SIGMOD Rec.* 50, 30–35. doi: 10.1145/3516431.3516438
- Cohoon, J. M., Nigai, S., and Kaye, J. J. (2011). Gender and computing conference papers. *Commun. ACM* 54, 72–80. doi: 10.1145/1978542.1978561
- Deryugina, T., Shurchkov, O., and Stearns, J. (2021). Covid-19 disruptions disproportionately affect female academics. *AEA Papers Proc.* 111, 164–168. doi: 10.1257/pandp.20211017
- Eagly, A. H. (2018). “Have gender stereotypes changed? yes and no,” in *Presentation at INSEAD Women at Work Conference, February, Vol. 17*. Available online at: <https://www.youtube.com/watch?v=ewOsOtHB-18> (accessed June 25, 2022).
- Ertl, B., Luttenberger, S., and Paechter, M. (2017). The impact of gender stereotypes on the self-concept of female students in stem subjects with an under-representation of females. *Front. Psychol.* 8, 703. doi: 10.3389/fpsyg.2017.00703
- Frachtenberg, E., and Kaner, R. D. (2021). “Representation of women in hpc conferences,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (St. Louis, MO), 1–14.
- Frachtenberg, E., and Kaner, R. D. (2022). Underrepresentation of women in computer systems research. *PLoS ONE* 17, e0266439. doi: 10.1371/journal.pone.0266439
- Holman, L., Stuart-Fox, D., and Hauser, C. E. (2018). The gender gap in science: how long until women are equally represented? *PLoS Biol.* 16, e2004956. doi: 10.1371/journal.pbio.2004956
- Huang, J., Gates, A. J., Sinatra, R., and Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proc. Natl. Acad. Sci. U.S.A.* 117, 4609–4616. doi: 10.1073/pnas.1914221117
- Inno, L., Rotundi, A., and Piccialli, A. (2020). Covid-19 lockdown effects on gender inequality. *Nat. Astron.* 4, 1114–1114. doi: 10.1038/s41550-020-01258-z
- Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., and Strohmaier, M. (2016). “Inferring gender from names on the web: a comparative evaluation of gender detection methods,” in *Proceedings of the 25th International conference companion on World Wide Web* (Montreal, QC), 53–54.
- Martin, J. L. (2014). Ten simple rules to achieve conference speaker gender balance. *PLoS Comput. Biol.* 10, e1003903. doi: 10.1371/journal.pcbi.1003903
- Mattauch, S., Lohmann, K., Hannig, F., Lohmann, D., and Teich, J. (2020). A bibliometric approach for detecting the gender gap in computer science. *Commun. ACM* 63, 74–80. doi: 10.1145/3376901
- Misra, J., Lundquist, J. H., and Templer, A. (2012). Gender, work time, and care responsibilities among faculty 1. *Sociol. Forum* 27, 300–323. doi: 10.1111/j.1573-7861.2012.01319.x
- Piatek-Jimenez, K., Cribbs, J., and Gill, N. (2018). College students’ perceptions of gender stereotypes: making connections to the underrepresentation of women in stem fields. *Int. J. Sci. Educ.* 40, 1432–1454. doi: 10.1080/09500693.2018.1482027
- Roper, R. L. (2019). Does gender bias still affect women in science? *Microbiol. Mol. Biol. Rev.* 83, e00018–19. doi: 10.1128/MMBR.00018-19
- Thébaud, S., and Charles, M. (2018). Segregation, stereotypes, and stem. *Soc. Sci.* 7, 111. doi: 10.3390/socsci7070111
- Tomkins, A., Zhang, M., and Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. *Proc. Natl. Acad. Sci. U.S.A.* 114, 12708–12713. doi: 10.1073/pnas.1707323114
- Wang, L. L., Stanovsky, G., Weihs, L., and Etzioni, O. (2021). Gender trends in computer science authorship. *Commun. ACM* 64, 78–84. doi: 10.1145/3430803

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

# Frontiers in Computer Science

Explores fundamental and applied computer science to advance our understanding of the digital era

An innovative journal that fosters interdisciplinary research within computational sciences and explores the application of computer science in other research domains.

## Discover the latest Research Topics

[See more →](#)

### Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne, Switzerland  
[frontiersin.org](https://frontiersin.org)

### Contact us

+41 (0)21 510 17 00  
[frontiersin.org/about/contact](https://frontiersin.org/about/contact)

