frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 20 March 2017
doi: 10.3389/frobt.2017.00008

A Human-Centered Approach to One-Shot Gesture Learning

*Maria Eugenia Cabrera and Juan Pablo Wachs**

Intelligent Systems and Assistive Technologies (ISAT) Laboratory, School of Industrial Engineering, Purdue University, West Lafayette, IN, USA

This article discusses the problem of one-shot gesture recognition using a human-centered approach and its potential application to fields such as human–robot interaction where the user's intentions are indicated through spontaneous gesturing (one shot). Casual users have limited time to learn the gestures interface, which makes one-shot recognition an attractive alternative to interface customization. In the aim of natural interaction with machines, a framework must be developed to include the ability of humans to understand gestures from a single observation. Previous approaches to one-shot gesture recognition have relied heavily on statistical and data-mining-based solutions and have ignored the mechanisms that are used by humans to perceive and execute gestures and that can provide valuable context information. This omission has led to suboptimal solutions. The focus of this study is on the process that leads to the realization of a gesture, rather than on the gesture itself. In this case, context involves the way in which humans produce gestures—the kinematic and anthropometric characteristics. In the method presented here, the strategy is to generate a data set of realistic samples based on features extracted from a single gesture sample. These features, called the "gist of a gesture," are considered to represent what humans remember when seeing a gesture and, later, the cognitive process involved when trying to replicate it. By adding meaningful variability to these features, a large training data set is created while preserving the fundamental structure of the original gesture. The availability of a large data set of realistic samples allows the use of training classifiers for future recognition. The performance of the method is evaluated using different lexicons, and its efficiency is compared with that of traditional *N*-shot learning approaches. The strength of the approach is further illustrated through human and machine recognition of gestures performed by a dual-arm robotic platform.

Keywords: gesture recognition, one-shot learning, embodiment, robotics, human–computer interaction

OPEN ACCESS

Edited by:
Ka-Wai Kwok,
University of Hong Kong, Hong Kong

Reviewed by:
Sigal Berman,
Ben-Gurion University of the Negev Beersheba, Israel
Jindong Liu,
Imperial College London, UK
Kaiyu Hang,
KTH Royal Institute of Technology, Sweden

****Correspondence:***
Juan Pablo Wachs
jpwachs@purdue.edu

Specialty section:
This article was submitted to Biomedical Robotics, a section of the journal Frontiers in Robotics and AI

Received: *18 August 2016*
Accepted: *20 February 2017*
Published: *20 March 2017*

Citation:
Cabrera ME and Wachs JP (2017) A Human-Centered Approach to One-Shot Gesture Learning. Front. Robot. AI 4:8. doi: 10.3389/frobt.2017.00008

INTRODUCTION

Gestures are a key component of human–human interactions (Kendon, 1986). Therefore, we expect machines and service robots to be able to understand this form of interaction as intuitively as humans do. Having seen a gesture only once, we are then able to recognize it the next time it is presented because of our capability to learn from just a few examples and to make associations between concepts (Brown and Kane, 1988; Ormrod and Davis, 2004; Johnson et al., 2005). Modeling this capability is one of the main challenges faced in the development of natural human–robot interaction (HRI). Currently, a number of learning sessions must take place before machines can be used in a natural and straightforward setting (Adams, 2005).

The problem of recognizing gestures from a single observation is called one-shot gesture recognition (Escalante et al., 2017). In the aim of providing natural interaction with machines, a framework must be developed to include the adaptability that current approaches lack. The limited amount of information provided by a single observation makes this an ill-posed problem if an approach based exclusively on data mining or statistics is adopted; instead, some form of context is required.

Previous work in this area has relied on computer vision techniques, extracting motion and orientation descriptors to train and further classify gestures based on a single training instance (Wu et al., 2012; Konecny and Hagara, 2014). However, gesture recognition involves intrinsic difficulty discerning between motions consistently repeated across examples of the same class, as well as having to cope with the high variability of human actions. When only one example is provided, this task becomes even more challenging, increasing the risk of overfitting as well as imposing limits on generalizability (Fe-Fei et al., 2003).

By including the human aspect within the framework, the kinematic and psychophysical attributes of the gesture production process can be used to support recognition. This approach involves a strategy in which these attributes are relied upon to generate a data set of realistic samples based on a single example and is therefore within the scope of one-shot learning. Using a single-labeled example, multiple instances of the same class are generated synthetically, augmenting the data set and enabling one-shot learning.

The recognition problem is based on using the generation process for a gesture instance rather than the instance itself. The proposed method is able to capture significant variability while maintaining the fundamental structure of the gesture, thereby accounting for the stochastic aspects of gesture production that are associated with the inherent non-linearity of human motor control. The extraction of the "gist of the gesture" consists in finding salient characteristics in the given gesture example that transcend human variability and are present in all examples of the same gesture class.

This study is a continuation of the research published by Cabrera and Wachs (2016). It is expanded by incrementing the number of data sets used for training and testing; an additional classifier is implemented and tested; new results are presented regarding the robotic implementation; and a new metric is proposed to compare the presented approach for one-Shot learning in terms of efficiency against its N-shot counterpart.

Background

Gestures are an intrinsic part of human communication, either complementing spoken language or replacing it altogether (Hewes, 1992). A gesture can be defined as a deliberate set of motions executed with any body part to convey a message or evoke an action (McNeill and Levy, 1980; Kendon, 1990). The scope of this review encompasses gestures performed with the upper limbs; they may be static (i.e., a pose) or dynamic. Given the relevance of gestures as means of human–human and human–machine interaction, they have been the subject of research in linguistics, computer science, engineering, and cognitive sciences. Researchers have been studying how gestures are produced, perceived, and mimicked, as well as how computer systems can detect and recognize them. This last area is especially relevant to human–computer interaction (Pavlovic et al., 1997; Rautaray and Agrawal, 2015), HRI (Nickel and Stiefelhagen, 2007; Yang et al., 2007), and assistive technologies (Jacob and Wachs, 2014; Jiang et al., 2016b), where humans rely on accurate recognition by machines.

The scope of the gesture recognition problem ranges from *N*-shot learning, in which several observations have been presented to the machine at earlier points in time, through to zero-shot learning (Palatucci et al., 2009; Socher et al., 2013), in which no observations have yet been made. Within this range lies the case of single-instance recognition (one-shot learning), in which only a single sample has been observed previously, and it is on this problem of one-shot learning as applied to gesture recognition that this paper focuses.

Gesture Communication

The phenomenon of gesturing is found across cultures, ages, and tasks (Goldin-Meadow, 1999). There are many ways in which gestures can be interpreted and classified, according to how they are used for communication. For example, in language, gestures are considered a natural way to communicate. Such gestures can be used to accompany speech or substitute for it entirely. The use of head, body, and hand gestures, including gaze, in a conversation conveys information about an individual's internal mental state, intention, and emotion (Kendon, 1980). This is also true when verbal expression is not part of the communication, for example, sign language for people with hearing disabilities (Bellugi, 1979). Furthermore, some gestures are not used as part of explicit communication, but rather arise as artifacts reflecting an individual's internal state of mind or a physical state (Kendon, 1994).

In one of the most widely used classifications of gestures, McNeill (1992) has identified four major types of gestures used by speakers: iconic, metaphoric, deictic, and beat gestures. Although iconic gestures capture semantic aspects of speech content, metaphoric gestures can provide a pictorial or abstract representation of speech. Deictic or pointing gestures can refer to both space and time, and beat gestures are usually cyclic to represent a rhythmic pulsation, similar to musical patterns.

There are, however, other gestures that are performed spontaneously during speech and that do not fit within this classification. These gestures are involved in aiding the thought process of the speaker (Krauss et al., 1991), rather than helping a listener understand an idea (Kendon, 1994).

N-Shot Learning

This is the most common method used for learning and pattern recognition. The key idea is that classifiers are trained using multiple observations. Although the number of observations may be of the order of 10 (Yamato et al., 1992; Hertz et al., 2006; Wasikowski and Chen, 2010), it is more common for hundreds of observations to be made (Rigoll et al., 1997; Liang and Ouhyoung, 1998; Wei et al., 2011; Jost et al., 2015; Mapari and Kharat, 2015) and sometimes even thousands (Babu, 2016; Sun et al., 2015; Zheng et al., 2015; Zhou et al., 2015). The number depends strongly on the application, which may vary from object or face recognition in images or clips (Serre et al., 2005; Huang et al., 2007; Toshev et al., 2009) to gestures or patterns coming from complex multimodal inputs (Jaimes and Sebe, 2007; Escalera et al., 2016). Some of the major challenges regarding recognition lie in representation, learning, and detection (Lee et al., 2016). The models used to describe gestures involve empirical and tunable parameters and a variety of descriptors to encompass the complexity and diversity among gestures. At the same time, models need to be flexible enough to handle high variability within the same class of gestures, while discriminating between classes (Caulfield and Heidary, 2005; Marron et al., 2007; Parikh and Grauman, 2011). Usually, variability is introduced into the model by using several training examples along with techniques based on prior information (Zaffalon and Hutter, 2002; Wang et al., 2015; Sarup et al., 2016).

Gesture recognition algorithms differ in many aspects. An initial classification may be done with respect to the overall structure of the adopted framework, i.e., the way in which the recognition problem is modeled. In particular, some approaches are based on machine learning techniques, where each action is described as a complex structure (Ikizler and Duygulu, 2009; Merler et al., 2012), whereas others involve simpler representations where a hand pose is represented as signatures and moments of inertia (Albrecht et al., 2003; Bhuyan et al., 2011).

One-Shot Learning

The one-shot learning paradigm relies on the use of a single training instance to classify future examples of the same image, gesture, or class. Most of the work reported in the literature on this type of learning paradigm, particularly in the context of gesture recognition, is based on appearance models and feature extraction techniques. There are techniques that are on the boundary between one-shot and N-shot (usually with $N < 5$ samples), and although these are not referred to directly as one shot, their capability for generalization makes them especially attractive in the field of one-shot learning (Rekabdar et al., 2015).

Some of the seminal work on one-shot learning has been based on a Bayesian framework that considers both the shape and appearance of objects to be classified in different images (Fe-Fei et al., 2003); by using a probabilistic framework, the density of feature locations provides scale and translational invariance. Once several categories have been learnt using this framework, a new category can be learnt using a single image as training (Fei-Fei et al., 2006); the reported accuracy was 82% in that case.

Lake et al. (2011) used a probabilistic approach with Markov Chain Monte Carlo and Metropolis–Hastings algorithms to detect the order of strokes in characters, comparing between different models and humans. The classification accuracy reported was 62%. Maas and Kemp (2009) used Bayes Nets on a public data set from Ellis Island to solve the "Randeria problem," which attempts to match the attributes of immigrants to their country of origin. To adjust to a one-shot learning framework, they removed all but one set of training examples each of which was based on a single category and then used the remaining example as testing data; the average accuracy reported among attributes was 70%.

An important landmark in one-shot learning applied to gestures was the ChaLearn Looking at People Challenge initiated by Microsoft in 2011 (ChaLearn Looking at People, 2014). This challenge involves a competition to design a one-shot gesture recognition method using Microsoft's Kinect technology (Zhang, 2012). Kinect was used because of its ability to gather color and depth information from video streams. For 2 years, a vast data set (CGD11), of both development and validation batches, was used worldwide as training and testing data in the competition; the results for both years were reported by Guyon et al. (2012, 2013) with partial success being achieved. Among the results reported, the Levenshtein distance (LD) (Levenshtein, 1966) was between 0.15 and 0.3 (the ideal distance is represented by 0 and the worst by 1). A common theme of the proposed methods was an emphasis on gesture representation as strictly machine learning and classification of observations regardless of the process involved in their generation. Most selected features were selected to portray appearance and motion using color or depth frames in the video input. There was no mention of the relevance of the shape or other characteristics of the human body performing the gestures.

Wan et al. (2013) extended the scale-invariant feature transform to spatio-temporal feature descriptors, known as a "bag of visual words" to build a codebook. Testing videos were then processed, and the codebook was applied to further classify using a K-nearest-neighbors algorithm. Their LD reached 0.18. Fanello et al. (2013) applied adaptive sparse coding to capture high-level feature patterns based on a three-dimensional histogram of flow and a global histogram of oriented gradient, classified by a linear support vector machine (SVM) using a sliding window and reported an LD of 0.25. Wu et al. (2012) utilized both RGB color and depth information from Kinect and adopted an extended motion history image representation as the motion descriptor and a maximum correlation coefficient as the discriminatory method. They found an LD of 0.26. Konecny and Hagara (2014) took a different approach, using histogram of oriented gradients to describe the visual appearance of gestures, with dynamic time warping (DTW) as the classification method. Their LD was 0.17.

More recently, Escalante et al. (2017) have described a method in which a two-dimensional map of motion energy is obtained for each pair of consecutive frames in a video and then used for recognition after applying principal component analysis. One-shot learning has also been applied in scene location (Kwitt et al., 2016), grasping of novel objects (Kopicki et al., 2016), and facial expression recognition (Jiang et al., 2016a).

MATERIALS AND METHODS

This section presents details of the implementation of a method to achieve one-shot gesture recognition through the "gist of the gesture." An overview of the implemented system is shown in **Figure 1**. Initially, the system requires a labeled example from a user. This is achieved in the following way. First, a gesture is performed by a user and is detected and recorded using a Kinect sensor. Using the skeleton data provided by the sensor, salient characteristics are extracted, which we refer to as the *gist of the gesture*, and are used to recreate new realistic observations, which resemble that provided by the user (Cabrera and Wachs, 2016). This process is repeated until a large data set of observations is generated. This data set constitutes the training set of an arbitrary

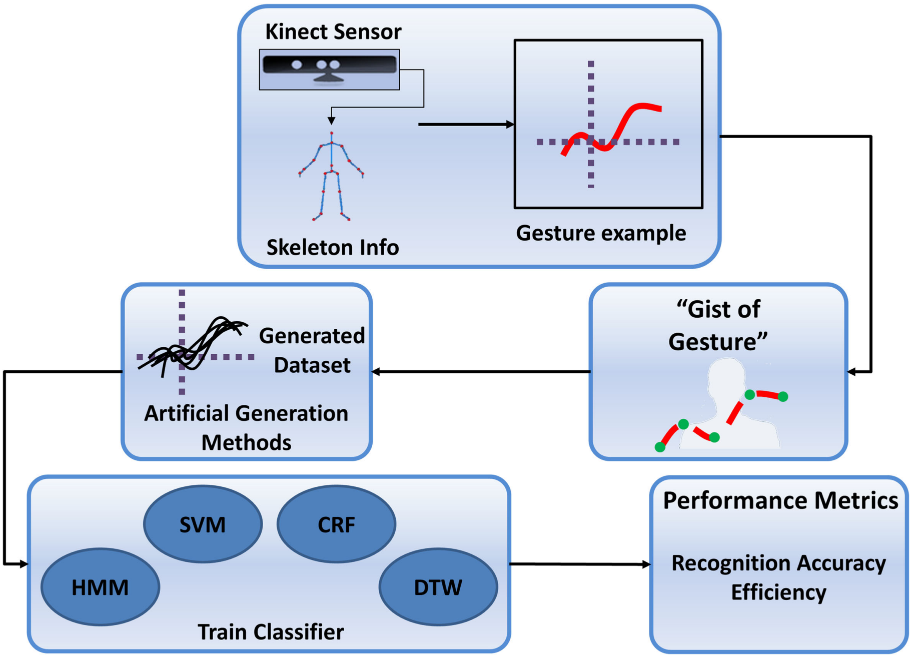


FIGURE 1 | System overview.

classifier. Metrics include independence from classification strategy and efficiency compared with N-shot learning. To further demonstrate the strength of the approach, gestures performed using a dual-arm robotic platform are detected and recognized using the previously trained classifiers.

The motivation behind the gist of the gesture is to understand how humans gesture and what determines the forms of the gestures produced. Furthermore, cognitive signatures related to observed gestures may be used to "compress" a gesture in memory while retaining its intrinsic characteristics. When a gesture is recalled, these key points associated with the cognitive signatures are used to "decompress" the gesture into a physical expression. In addition, multiple instances of the same gesture will share key common motion components among all instances, regardless of the variability associated with human performance. The fact that positive correlations have been observed between abrupt changes in motion and spikes in electroencephalographic signals associated with the motor cortex supports the hypothesis of a link between inflection points (IPs) in motion and cognitive processes.

In a preliminary experiment, it was found a relationship between the timing of *mu* oscillations and kinematic IPs, such that IPs were followed by interruptions in *mu* suppression approximately 300 ms later. This lag is consistent with the notion that IPs may be utilized as place holders involved in conscious gesture categorization. This is the first evidence (from cognitive, objective, and empirical studies) that these specific landmarks represent stronger footprints in people's memory than other points (Cabrera et al., 2017a).

Problem Definition

Let $\mathcal{L}$ describe a set or "lexicon" formed by N gesture classes $\mathcal{G}_i$, $\mathcal{L} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots \mathcal{G}_i, \ldots, \mathcal{G}_N\}$. Each gesture class is expressed through its realization by a set of gesture instances g_k^i. In a way, the gesture class is a prototype group, and the members of that group are the instances $g_k^i \in \mathcal{G}_i$, where $k = 1, \ldots, M$ is the number of instances of gesture class i. Each gesture instance is a concatenation of trajectory points in three dimensions, $g_k^i = \{(x_1, y_1, z_1), \ldots, (x_h, y_h, z_h)\}$, where h is the total number of points within that gesture instance.

Using one instance per class g_1^i, a set of place holders or IPs x_q^i, where $q = 1,\ldots,l$ and $l < h$, are computed. We refer to the set of values $\widetilde{G}_i$ as the "gist of a gesture" of gesture class i in lexicon $\mathcal{L}$:

$$\widetilde{G}_i = \left\{ x_q^i = (x_q, y_q, z_q) : x_q^i \in g_k^i, q = 1, \ldots, l, l < h \right\}.$$
$$\widetilde{G}_i \in \widetilde{G_{\mathcal{L}}}, i = 1, \ldots, N. \tag{1}$$

This set of values is obtained using a function $\mathcal{M}$ (2) that maps from the gesture dimension h to a reduced dimension l:

$$\widetilde{G}_i = \mathcal{M}\left(g_k^i\right), k = 1, i = 1, \ldots, N; g_k^i \in \mathbb{R}^h; \widetilde{G}_i \in \mathbb{R}^l; l < h. \tag{2}$$

This compact representation is then used to generate artificial gesture examples $\hat{g}_k^i$ for each $\mathcal{G}_i$. This is done through a function $\mathcal{A}$ (3) that maps from dimension h to gesture dimension l using the forward approach explained in Section "Artificial Observation Generation":

$$\hat{g}_k^i = \mathcal{A}\left(\widetilde{G}_i\right), k = 1, \ldots, M; i = 1, \ldots, N. \tag{3}$$

A function Ψ (4) maps gesture instances to each gesture class using the artificial examples:

$$\Psi : \hat{g}_k^i \to \mathcal{G}_i. \tag{4}$$

Then, for future instances g^u of an unknown class, the problem of one-shot gesture recognition is defined as follows:

$$\text{Max } Z = \mathcal{W}\left\{\Psi(g^u), \mathcal{G}_i\right\}$$
$$\text{s.t.} i = 1, 2, \ldots, N, i \in \mathbb{Z}^+, \mathcal{G}_i = \Psi\left(g_1^i\right), \Psi\left(g^u\right) \in \mathcal{L}, \quad (5)$$

where $\mathcal{W}$ is a metric function (e.g., accuracy).

Artificial Observation Generation

To generate artificial observations $\hat{G}^i$ using the forward method, the stored location for each IP x_q^i, where $q = 1, \ldots, l$, is used as the mean value μ for a mixture of Gaussians (6), while the quadrant information relative to the user's shoulder of all $x \in g_1^i$ of the hand's trajectory is used to estimate the variance.

Considering a vector of dimension d (in this case 3), this Gaussian mixture model (GMM) is fitted as follows:

$$(\hat{x}; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k, \pi_k) = \sum_{k=1}^{m} \pi_k p_k(x), \pi_k \geq 0, \sum_{k=1}^{m} \pi_k = 1,$$
$$p_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2} \sigma_k{}^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\sigma}_k{}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right\} \quad (6)$$

where m represents the mixtures in the model, p_k is the normal distribution density with mean μ_k taken as the location of each IP and a covariance matrix σ_k that is positive semidefinite, and π_k is the weight of the kth mixture, with all the weights summing to unity.

To estimate the variance, each point x in the sample trajectory is assigned to a quadrant with respect to the user's shoulder, $x_j = q(x)_c$, where $c = 1, 2, 3, 4$, using the reference frame shown in **Figure 2**. The next step uses the points x_j from each quadrant as samples to estimate the variance of each quadrant as

$$\sigma_c = \frac{1}{n_c - 1} \sum_{j=1}^{n_c} (x_j - \boldsymbol{\mu}_c)^2, x_j = q(x)_c \in \mathbb{R}^3, c = 1, 2, 3, 4, \quad (7)$$

which in turn adjusts the parameters for the generated GMM for each IP.

With different sets of IPs, generated using the GMM, and the curvature information related to the original gesture trajectory, artificial trajectories are generated. The points p_i and p_{i+1} are used along with the curvature c_i to generate smooth segments for all i in the set of IPs. The basic algorithm is outlined in **Algorithm 1**, taking as input a single labeled example from class s, g_1^s, (8) consisting of A points, acquired using the skeleton information from a Kinect:

$$g_1^s = \{(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_k, y_k, z_k), \ldots, (x_A, y_A, z_A)\}. \quad (8)$$

Performance Metrics

Once the gist of a gesture $\tilde{G}_i$ has been extracted from a single example g_1^i of each gesture class $\mathcal{G}_i$ within a lexicon $\mathcal{L}$ with $i = 1, \ldots, N$, and an artificially enlarged data set has been created from it as $\mathbb{G} = \{\hat{G}^1, \ldots, \hat{G}^i, \ldots, \hat{G}^N\}$, the goal is to evaluate the performance of the method in terms of generalization and recognition of future instances. The proposed method must work with different gestures and lexicons, as well as being independent from a chosen classification strategy. The accuracy metric $A\%$ is defined (9) as the ratio of the number of true hits to the total number of samples:

$$A\% = \frac{\text{total}_{\text{true-hits}}}{\text{total}_{\text{samples}}}. \quad (9)$$

The proposed method has to be generalizable to allow comparison with N-shot learning approaches and thereby empirically determine the number of samples required for each classifier to reach the same recognition accuracy obtained when training them with artificially generated samples. Therefore, the following metric of efficiency $\eta(\cdot)$ is proposed (10) to express the extent to which the presented approach is more efficient than the standard N-shot learning approach, given the number of samples

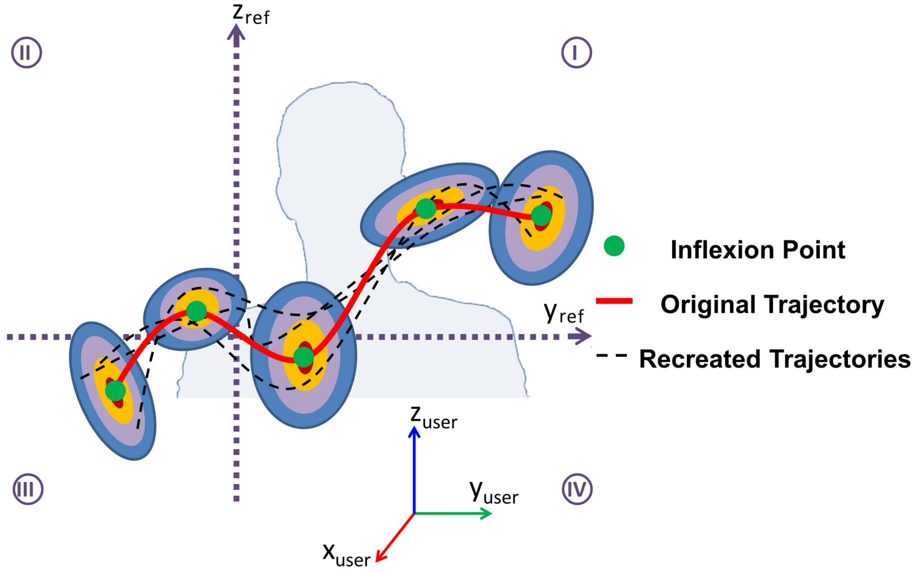


FIGURE 2 | Visual representation of how features corresponding to the "Gist of the Gesture" are used to create new observations.

ALGORITHM 1 | Generating artificial observations from a single sample.

Input:

g_1^s three-dimensional hand trajectory of a gesture instance of class s
$\mathbf{x}_s = (x_s, y_s, z_s)$ three-dimensional position of the shoulder
M number of artificial trajectories to generate

```
// Inflection points (IPs)
  $\mathbf{x}_{IP} \leftarrow \mathbf{x}_i \in g_1^s,\ \frac{d\mathbf{x}^2}{d^2t}\Big|_{\mathbf{x}=\mathbf{x}_i} = 0, i = 1$
// Interval between IPs
  $I_j = \{\mathbf{x} \mid \mathbf{x} \in (\mathbf{x}_i, \mathbf{x}_{i+1})\}, j = 1,..., l-1$
  for $l-1$ iterations do
// Convexity for interval $I_j$
      $C_j = \mathrm{sign}\left(\frac{d\mathbf{x}^2}{d^2t}\right), \mathbf{x} \in I_j$
  end for
// Determine quadrant location based on $\mathbf{x}_s$
  if $y_i > y_s, z_i > z_s$
      $q(\mathbf{x}_i) = 1$
  elseif $y_i < y_s, z_i > z_s$
      $q(\mathbf{x}_i) = 2$
  elseif $y_i < y_s, z_i < z_s$
      $q(\mathbf{x}_i) = 3$
  else // $y_i > y_s, z_i < z_s$
      $q(\mathbf{x}_i) = 4$
  end if
  // Variance estimation
  $\sigma_c = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} (\mathbf{p}_i - \boldsymbol{\mu}_c)^2,\ \mathbf{p}_i = |q(\mathbf{x}_c)|_c \in \mathbb{R}^3$
  // Generate GMM
  $\Gamma_i = \sum \sim N(\mathbf{x}_i, \boldsymbol{\sigma}_c), i = 1,..., m, c = 1,...,4$
  for $M$ iterations do
  // Sample $\Gamma_i$ to obtain new IP $\mathbf{x}_i^*$
      $\mathbf{x}_{IP}^* \leftarrow \mathbf{x}_i^* \in \Gamma_i,\ i = 1, \ldots, m$
      for $l-1$ iterations do
  // Smoothly connect new IP
          $a_l \leftarrow \cup\ \mathrm{arc}\ (x_i^*, x_{i+1}^*, C_i)$
      end for
  end for
```

Output:

$\hat{G}^s = \{\hat{g}_1^s,\ \hat{g}_2^s, \ldots, \hat{g}_K^s\}$ set of artificial trajectories for gesture class s

($\text{samples}_{\text{cutoff}}$) required to reach the same baseline for accurate recognition:

$$\eta = \frac{\text{samples}_{\text{cutoff}} - 1}{\text{samples}_{\text{cutoff}}}. \tag{10}$$

With a set of artificially generated observations for each gesture in a lexicon with its corresponding gesture class as label, each classification algorithm is trained and tested. Initially, testing data correspond to different samples from the acquired data set in the form of video inputs. Further details about data are provided in Section "Results."

Receiver operating characteristic (ROC) curves were created to indicate performance. Each ROC curve was drawn using different thresholds to assign the predicted class label. Different values across the ROC curve show the trade-off between recognition of true positives (hit rate) and false positives (false-alarm rate). The optimal recognition system would have a 100% recognition accuracy and no false positives. This maps to the point on the curve with the least distance to the top left intersection (1, 0), which is used to determine the best operating point for each classifier. The overall accuracy was determined by calculating the area under the curve.

Confusion matrices were obtained to further analyze the correlation between the actual and predicted labels of the testing data for each gesture class.

Implementation Details

Three lexicons were used to create artificial data sets from N gesture classes $\mathbb{G} = \{\hat{G}^1,\ \ldots,\ \hat{G}^i,\ \ldots, \hat{G}^N\}$ from one example g_1^i of each gesture class $\mathcal{G}_i$. One of the lexicons was customized to be used as an image manipulation (IMD) data set to interact with displayed images in a touchless manner. The two other lexicons are publicly available data sets, one of them related to the gesture challenge competition ChaLearn 2013 (GCD13) (ChaLearn Looking at People, 2014), with up to 20 different gestures related to Italian culture. The third data set was MSCR-12 from Microsoft Research (Fothergill et al., 2012), which included iconic and metaphoric gestures related to gaming and music player interaction.

The approach proposed in this article is independent of the specific form of classification. Furthermore, it is not conceived with any specific classification approach in mind. The expectation is that state-of-the-art classifiers will be selected to be trained with the artificial data sets created. This idiosyncratic approach was tested by training four different classification methods, currently used in state-of-the-art N-shot gesture recognition approaches, and adapting them to be used in one-shot gesture recognition.

Classification Algorithms

Four different classification algorithms were considered, and their performances were compared using the artificially generated data sets. The selected classification algorithms, namely, hidden Markov models (HMM), SVMs, conditional random fields (CRF), and DTW, are commonly used in state-of-the-art gesture recognition approaches. In the case of HMM and SVM, a one-versus-all scheme was used, while CRF and DTW provided a metric of likelihood for the predicted result after training was completed.

Each HMM comprised five states in a left-to-right configuration and was trained using the Baum–Welch algorithm, which has been shown to generate promising results in hand gesture recognition (Jacob and Wachs, 2014). An observation was classified based on the specific HMM chain that best explained that observation, that is, by determining which of the trained HMM outputs had the highest probability for a state sequence $\vec{z}$ given a new observation g^u and its intrinsic parameters (the initial state distribution vector π, the transition matrix A, and the observation probability within each state B), and thereby assigning the corresponding label to the new sample. This procedure can be expressed as follows:

$$\arg\max_i \{\log(P_i(\vec{z}\,|g^u; A_i\ , B_i, \pi_i)\},\ i = 1, \ldots, N. \tag{11}$$

For the SVM, each classifier in the one-versus-all scheme was trained using the radial basis function kernel. The library available in MATLAB® was used to implement SVM. In the case of CRF, the training examples were encoded using the BIO scheme to determine the beginning (B), inside (I), and outside (O) of a gesture. The CRF++ toolkit was used to train and test this classification algorithm (CRF++: Yet Another CRF Toolkit, 2016). The

DTW classification algorithm was implemented using the Gesture Recognition Toolkit (2016), which is a C++ machine learning library specifically designed for real-time gesture recognition.

Data Set 1: Image Manipulation (IMD)

This data set was custom designed to be used with a user interface that helps users manipulate displayed images using gestures. This lexicon comprised 11 gestures. The functions related to the gestures were zoom in/out, rotate clock/counterclockwise, pick/drop, previous/next, cut/paste, and erase. Some of the gestures with the associated hand trajectory and the extracted descriptors are shown in **Figure 3**.

This customized lexicon was developed for use in the medical telementoring application STAR, in which the mentor or expert is provided with multimodal interactions, including touch-based annotations, air gestures, and tool manipulation to guide a novice surgeon or trainee through a procedure (Andersen et al., 2016).

This data set was the smallest in terms of number of samples, but was the largest lexicon in terms of number of gesture classes within it. Six subjects performed each gesture class 5 times, for a total of 30 samples per gesture, using IRB-approved protocol #1609018129. Only depth and skeleton information was stored, thus keeping the recordings of each participant anonymous. The data set included 330 gesture clips. Eleven additional gesture clips were created, 1 for each gesture class, which were used to extract the IPs and create 200 artificial examples for each class. Those were used as the training set for the classifiers.

Data Set 2: ChaLearn Gesture Data Set 2013 (CDG13)

The second data set used in these preliminary results was a publicly available data set from the ChaLearn Gesture Competition 2013. The development data from this data set contained 7,754 gesture instances from a vocabulary of 20 gesture categories of Italian signs. These signs were cultural/anthropological gestures performed by different subjects. Examples of this lexicon with its extracted gist are shown in **Figure 4**.

From this data set, subsets of both gesture instances and classes were selected. The number of gesture classes in the lexicon was reduced to 10. This reduction was mainly due to the fact that this approach to gesture representation still does not include information about hand configuration. Without this, some of the gesture classes in the lexicon could be heavily confounded.

An additional issue related to the selection of CGD13 over CGD11 (the one-shot learning data set from 2011) is the available input information for each data set. While CGD11 provides gesture instances in color and depth format suited for one-shot gesture recognition, CGD13 offers additional input information, namely, audio and skeletal information from the subject. Given the nature of the proposed approach, based on anthropometric features such as the locations of the shoulder and hands, the use of skeletal information provides an advantageous starting point.

The reduced data set used for these preliminary results comprised 100 gesture instances of each gesture class for a total of 1,000 gesture instances, obtained from the development data of CGD13. Ten separate gesture instances, one from each gesture class, were

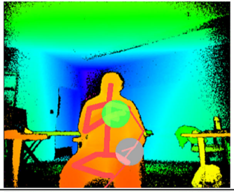
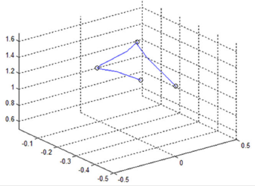
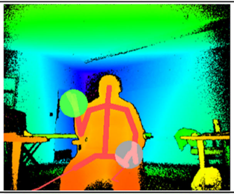
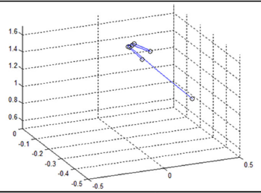
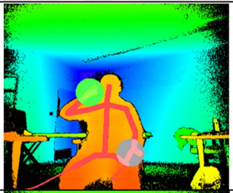
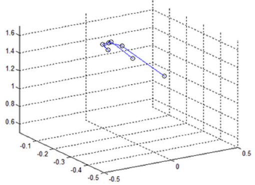
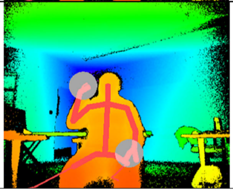
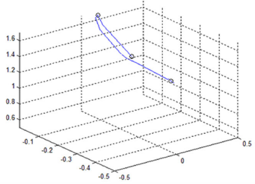
Gesture	Depth Image	Hand Trajectory	Feature Representation
"Next"			4 Inflexion points Quadrant Sequence: IV,I,II,IV Curvature: +, 0, +
"Rotate Clockwise"			7 Inflexion Points Quadrant Sequence: IV,II,II,I,I,II,II Curvature: +,+,+,-,+,+,+
"Rotate Counter-Clockwise"			7 Inflexion Points Quadrant Sequence: IV,I,I,II,II,II,I Curvature:+,+,+,+,+,+
"Pick"			3 Inflexion Points Quadrant Sequence: I,II,IV Curvature: -,+

FIGURE 3 | Examples of gestures from data set 1 (IMD).

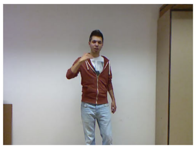
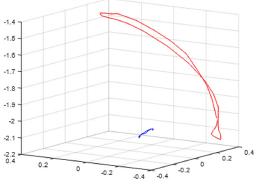
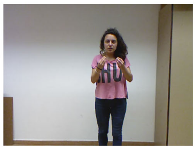
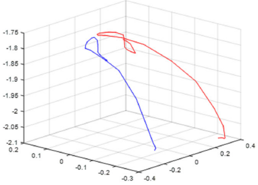
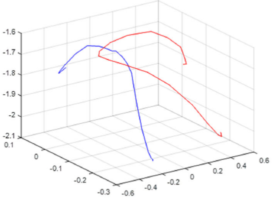
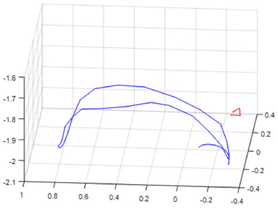
Gesture	Color Image	Hand Trajectory	Feature Representation
"Vieniqui" G2			3 Inflexion points Quadrant Sequence: III,II,III Curvature: +,+
"Chevuoi" G6			5 Inflexion Points Quadrant Sequence: III,II,IV,II,III Curvature: +,+,+,+
"Basta" G13			4 Inflexion Points Quadrant Sequence: III,IV,III,III Curvature: +,– ,+
"Tantotempo" G17			5 Inflexion Points Quadrant Sequence: IV,II,II,II,III Curvature: +,+,+,+

FIGURE 4 | Examples of gestures from data set 2 (CGD13).

used to extract the gist of the gesture. The gist was used to create 200 artificial examples for each class and assemble the training data set for the classifiers.

Data Set 3: Microsoft Research Cambridge-12 (MSRC-12)

This data set consisted of sequences of human movements, representing 12 different iconic and metaphoric gestures related to gaming commands and interaction with a media player. The data set included 6,244 gesture instances collected from 30 people. The files contained tracks of 20 joints estimated using the Kinect pose estimation pipeline.

A subset of this data set was selected. The number of gesture classes in the lexicon was reduced to 8. This reduction was related to the fact that some of the gesture classes performed whole-body motions such as kicking or taking a bow, whereas the proposed method focuses on gestures performed with the upper limbs.

From this data set, 100 gesture instances from each class were used as the testing set, for a total of 800 gesture motions. Eight additional instances, one for each gesture class, were used to extract the gist representation and from it create 200 artificial observations per class. This artificial data set was used as training data. Examples of the gestures in this lexicon are depicted in **Figure 5**.

Robotic Implementation

The use of a dual-arm robotic platform, the Baxter robot, to execute a testing set comprising artificially generated gestures provided a repeatable and accurate framework to test the motions of the generated examples. The variability among the examples did not come from the robotic performance, but the method used to generate the gesture instances. In a broader research perspective, the use of a robotic platform to recognize the artificially generated gestures opens the possibility to study the coherency in recognition between humans and machines, alternating the roles of executing and recognizing the gestures (Cabrera et al., 2017b).

To execute the gestures using the Baxter robot, a registration and mapping process were conducted. It involved finding the transformation between the space where the trajectories were generated and the robot's operational space. This task was accomplished through singular value decomposition. The homogeneous transformation matrix was found through least squares solution given 12 different points collected in both spaces.

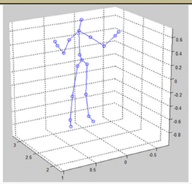
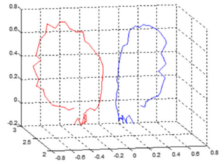
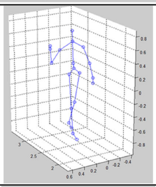
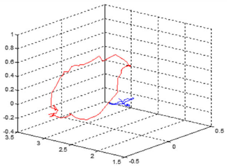
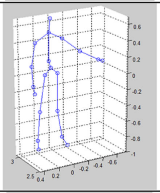
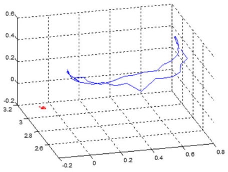
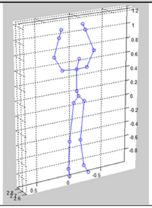
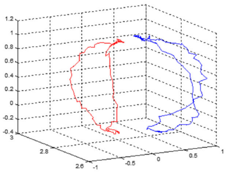
Gesture	Kinect Skeleton	Hand Trajectory	Feature Representation
"Wind Up"			5 Inflexion points Quadrant Sequence: IV,I,II,III,IV Curvature: +, +,+, +
"Throw"			4 Inflexion Points Quadrant Sequence: III,II,I,IV Curvature: +,+,-
"Next"			3 Inflexion Points Quadrant Sequence: IV,III,IV Curvature: +,+
"Start"			5 Inflexion Points Quadrant Sequence: III,III,II,III,IV Curvature: +,+,+,+

FIGURE 5 | Examples of gestures from data set 3 (MSRC-12).

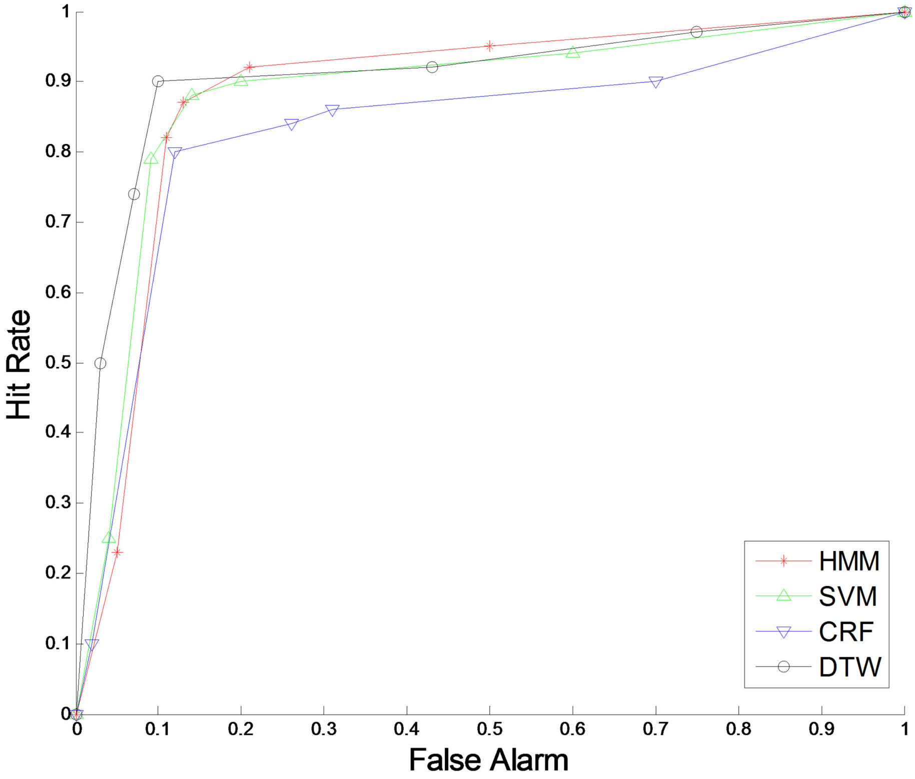


FIGURE 6 | Receiver operating characteristic curve for IMD data set.

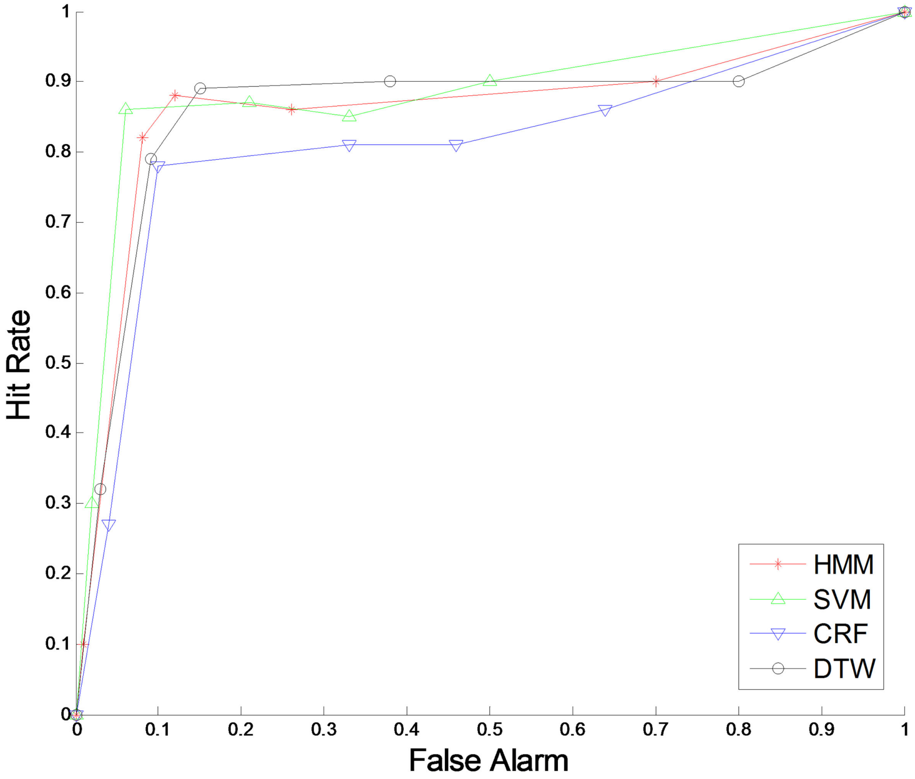


FIGURE 7 | Receiver operating characteristic curve for CGD13 data set.

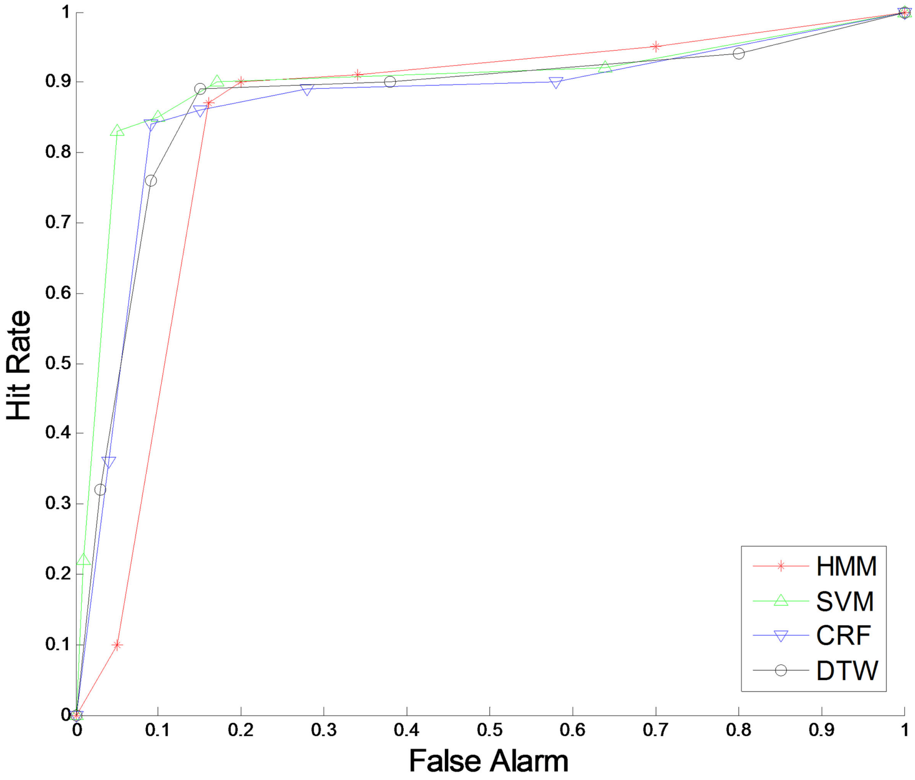


FIGURE 8 | Receiver operating characteristic curve for Microsoft Research Cambridge-12 data set.

A different approach to recognize Baxter's motions involves looking at the kinematics and reported end-effector's trajectories from the robot's robotic operating system (ROS.org | Powering the World's Robots, 2016) nodes. However, we preferred an approach that is agnostic with regard to the type of robot used to execute the gestures.

A simple computer vision method was developed to recognize the extremities of the performer and, through tracking, to estimate the trajectories constituting the gestures. Baxter's gestures were detected using the following procedure: (i) add a marker to the robot end effectors; (ii) segment the color using thresholding on the RGB channels of the view; (iii) apply morphological operators to obtain the candidate hand regions represented by blobs; and (iv) determine the center of mass for each blob (x_i, y_i) and then complement the three-dimensional representation using the depth value at that same center of mass in pixel values. These coordinates represent the position of a hand at time i.

TABLE 1 | Overall accuracies for different classifiers with different data sets.

	Hidden Markov models	Support vector machines	Conditional random fields	Dynamic time warping
IMD	92.7%	92.3%	89.4%	93.6%
CGD13	91%	92.6%	86.4%	89.5%
Microsoft Research Cambridge-12	90.6%	93.3%	91.4%	91.1%

Underlined results show the maximum recognition found for each data set.

RESULTS

Once the lexicon had been selected, the gist of the gesture extracted, and the data set expanded with artificial observations, four different classifiers were trained with these data sets to achieve one-shot gesture recognition. The performance of these classifiers is demonstrated in the following subsections in terms of accuracy and efficiency when compared with traditional N-shot learning approaches. Recognition accuracy across classifiers was also investigated, using the gestures generated with a dual-arm robotic platform.

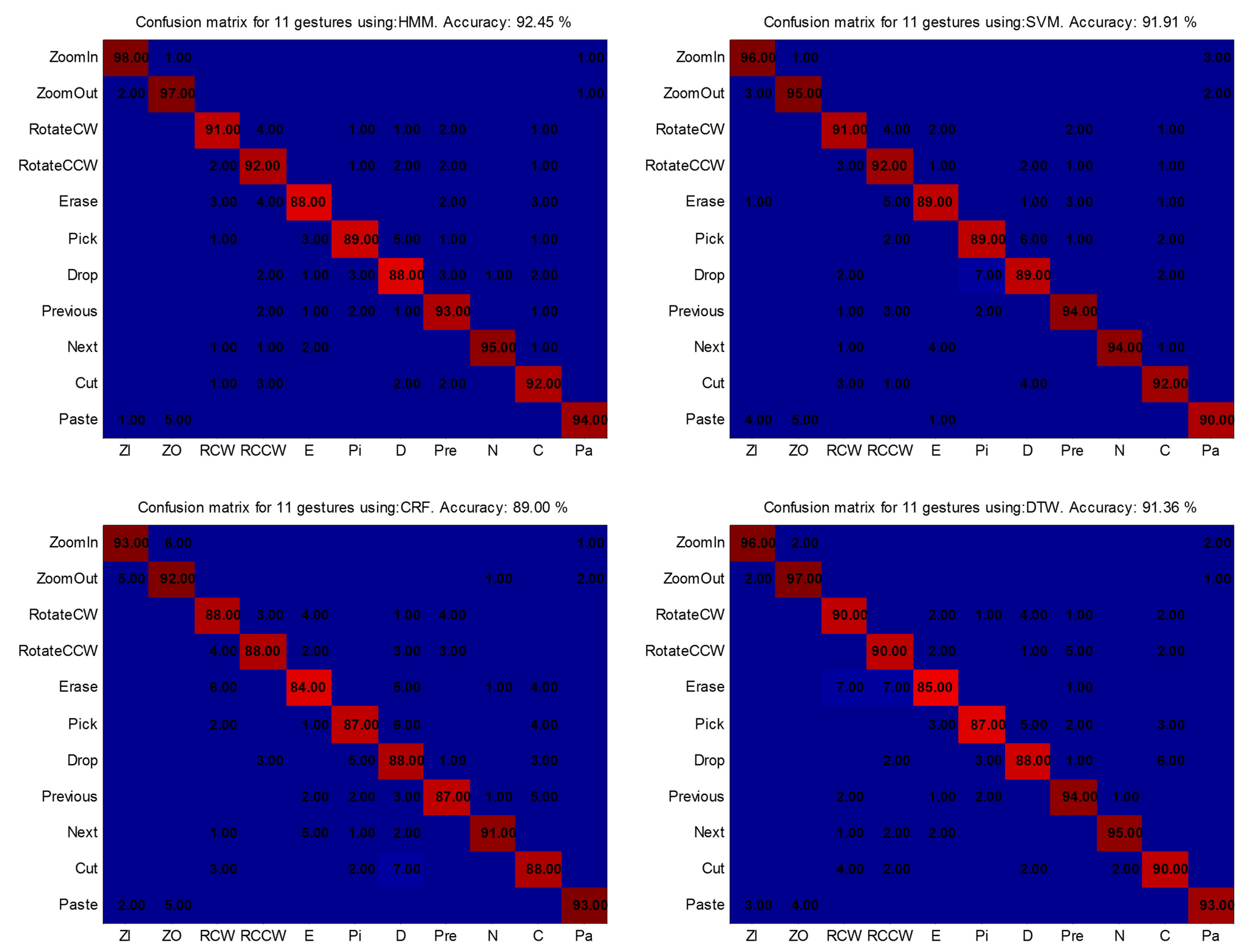


FIGURE 9 | Confusion matrices for the IMD data set. From upper left to lower right: hidden Markov models (HMM) (92.45%), support vector machines (SVM) (91.91%), conditional random fields (CRF) (89%), and dynamic time warping (DTW) (91.36%).

Accuracy

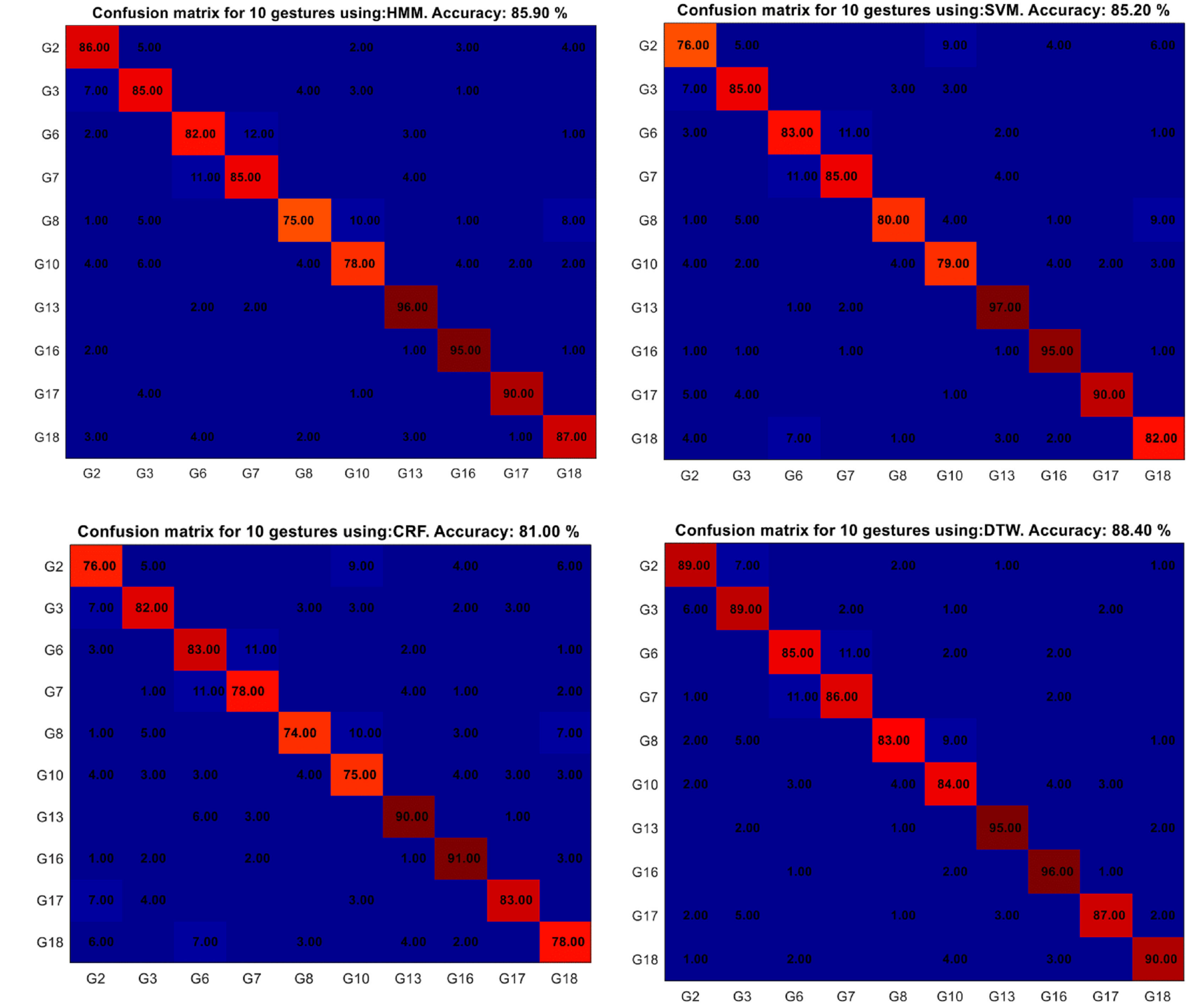
For each implemented classification algorithm, recognition accuracy as a metric was determined through ROC curves, one per data set, presenting all classifiers on the same graph. These are shown in **Figures 6–8**.

To obtain the ROC curves for each classifier, a free parameter was selected in each and varied to obtain different values for hit rate and false alarm. This free parameter was the likelihood of the predicted class for each classifier. The same parameter value was used four times for each point on the curve, dividing each data set and reshuffling, and keeping an even distribution of gesture examples between groups. The extreme points (0, 0) and (1, 1) were added to complete the curve range. Overall accuracies were obtained for all classifiers and are shown in **Table 1** for each data set.

Confusion matrices were used to determine the classifiers' performance. Within each matrix, the values reflect the proportion in which the predicted label assigned to a gesture instance matched its ground truth label. **Figures 9–11** show the results obtained with each data set.

The results obtained for the IMD data set were similar for all the classifiers when compared with the respective overall accuracy found through the ROC curves; specifically, HMM 92.45%, SVM 91.91%, CRF 89%, and DTW 91.36%. The gestures that were confounded the most included "Erase" and "Drop."

The results obtained for CGD13 were lower for all the classifiers when compared with their overall accuracy. This is related to the fact that all the classifiers had higher recognition rates at the expense of higher false-detection rates. By selecting the "best" classifier as the one closest to the goal (1, 0) on the ROC curve, lower recognition rates were found for each classifier. The recognition results were HMM 85.9%, SVM 85.2%, CRF 81%, and DTW 88.4%. The most confounded gestures across classifiers included G8 and G10.


Confusion matrix for 10 gestures using:HMM. Accuracy: 85.90 %

Confusion matrix for 10 gestures using:SVM. Accuracy: 85.20 %

Confusion matrix for 10 gestures using:CRF. Accuracy: 81.00 %

Confusion matrix for 10 gestures using:DTW. Accuracy: 88.40 %

FIGURE 10 | Confusion matrices for the CGD13 data set. From upper left to lower right: hidden Markov models (HMM) (85.9%), support vector machines (SVM) (85.2%), conditional random fields (CRF) (81%), and dynamic time warping (DTW) (88.4%).

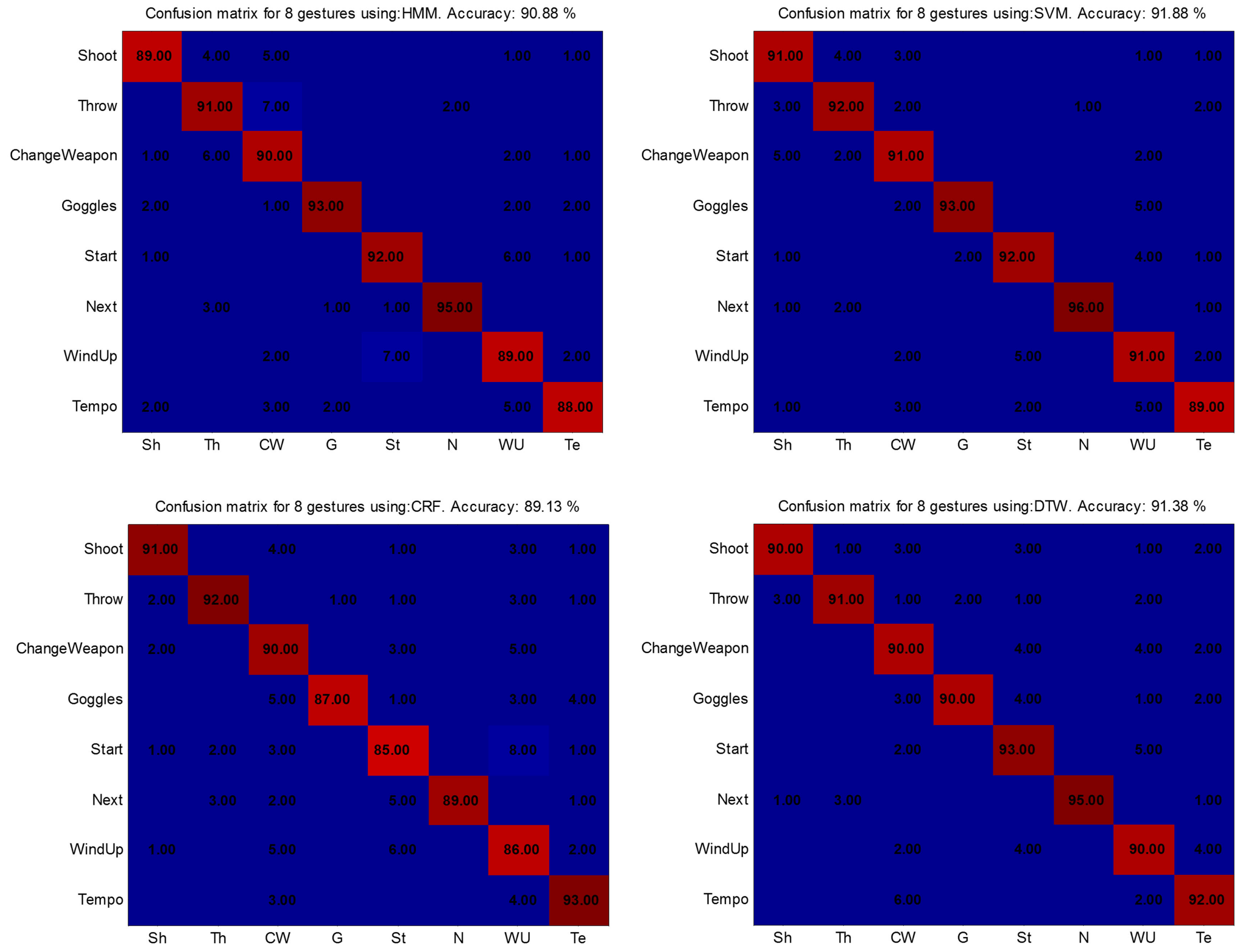


FIGURE 11 | Confusion matrices for the Microsoft Research Cambridge-12 data set. From upper left to lower right: hidden Markov models (HMM) (90.88%), support vector machines (SVM) (91.88%), conditional random fields (CRF) (89.13%), and dynamic time warping (DTW) (91.38%).

The results obtained for MSRC-12 were similar to those for GDC13 when compared against the overall accuracy for each classifier. The results with each classifier were HMM 90.88%, SVM 91.88%, CRF 89.13%, and DTW 91.38%. However, the recognition accuracy among classifiers was closer for this data set than for the other two.

Recognition Accuracy of Artificially Generated Gestures Performed by Baxter

A single data set was used to test this condition; MSCR-12 was selected. Twenty gesture instances per gesture class were executed by Baxter, detected, and transformed as new inputs for the classifiers. Confusion matrices were obtained for each classifier and are shown in **Figure 12**.

Recognition accuracy among all the classifiers was slightly lower than was found when testing the given data from the set; specifically, HMM 89.38%, SVM 90.63%, CRF 86.88%, and DTW 90%. However, the difference in number of samples also makes an impact. In addition, a noisy trajectory was more likely given that the detection approach was based solely on vision and did not use joint information directly available from Baxter.

Efficiency

The efficiency of the approach was compared with that obtained with N-shot learning. This comparison used the recognition accuracy obtained in Section "Accuracy" as a baseline to determine the number of samples required to achieve similar recognition results in a traditional N-shot learning approach.

Since the accuracy results obtained in Section "Accuracy" form the baseline for this new metric, new considerations are needed regarding the use of each data set. As a general criterion, data were divided, with 70% of each set being used for training and 30% kept for testing. The division was done after randomization, keeping the distribution between instances of each class equal.

Results for each combination **(data set–classifier)** are shown in **Figure 13**. For each graph, the green solid line represents the baseline accuracy obtained previously; the red dashed line represents the naïve accuracy, referring to the assignation of a random

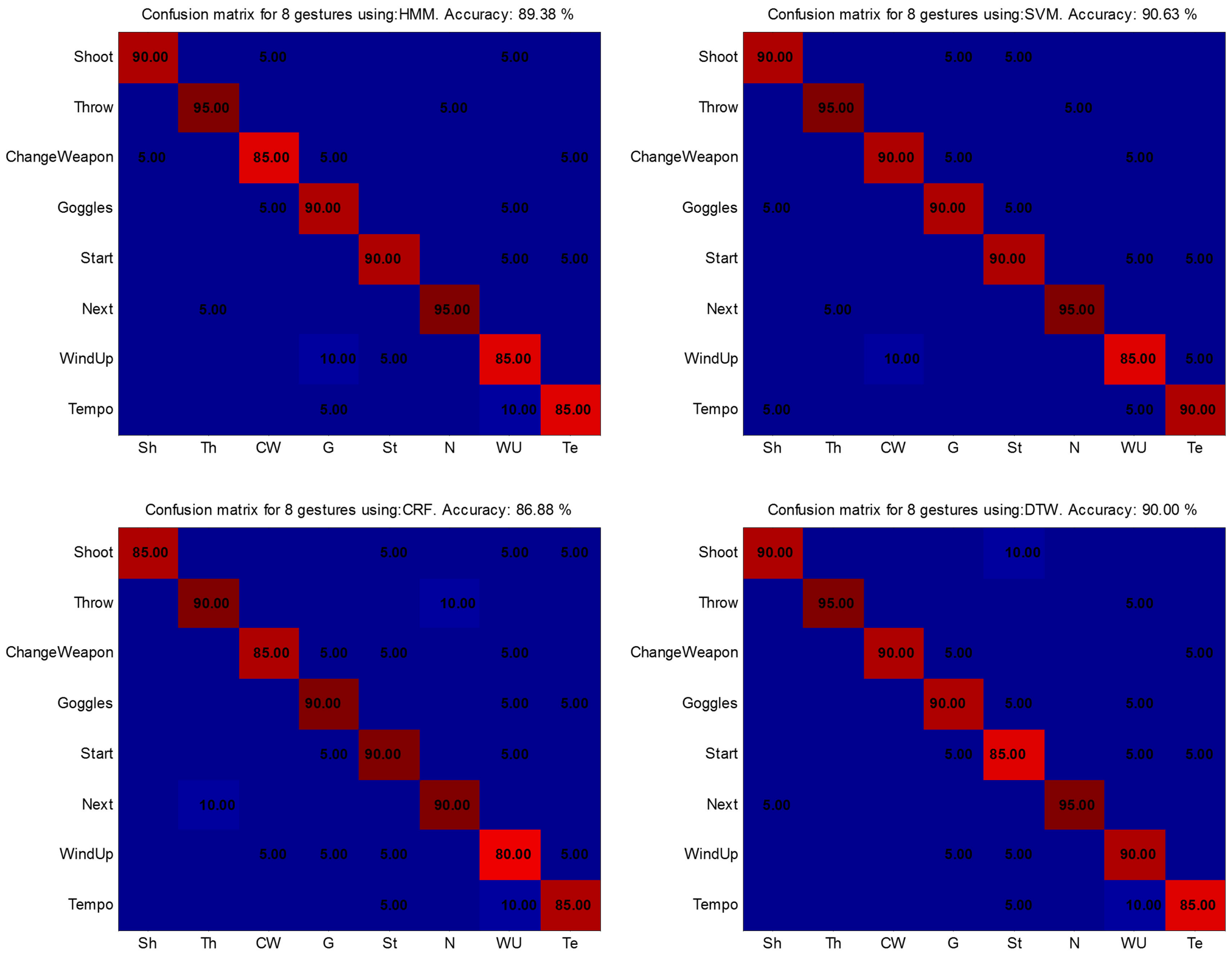


FIGURE 12 | Confusion matrices for the Microsoft Research Cambridge-12 data set performed by robotic platform Baxter. From upper left to lower right: hidden Markov models (HMM) (89.38%), support vector machines (SVM) (90.63%), conditional random fields (CRF) (86.88%), and dynamic time warping (DTW) (90%).

gesture class, with all the classes having equal prior probability. Finally, the blue line with dots represents the obtained recognition accuracy for a given percentage of training samples used.

The most obvious difference in performance between data sets occurs with IMD. This data set has fewer (30) samples per class compared with the others. Therefore, the efficiency is not as high. That is, using one observation instead of 100 gives a greater saving than using one observation instead of 30. No classifier was able to reach the accuracy baseline set by one-shot gesture recognition using the method proposed.

For CGD13, all classifiers needed more than 70% of the training data to reach the baseline recognition accuracy. In the case of the MSRC-12 data set, baseline accuracies were reached with 50% of the training data for SVM, while about 90% was required for CRF and DTW.

The cutoff values for the number of samples where the recognition accuracy for each classifier reached the baseline were used to determine the efficiency metric η. Results were only calculated for CGD13 and MSCR-12. Since the recognition accuracy never reached the baseline value for IMD, η tended to 1. These results are shown in **Table 2**.

DISCUSSION AND CONCLUSION

The obtained results show the performance of the method developed for one-shot gesture recognition through the gist of the gesture. They demonstrate the independence of the method with respect to the selected classification strategy. In addition, different gesture vocabularies were used: one customized lexicon for IMD interfaces and two different public data sets.

Previous results with the same ChaLearn data set were described by Escalera et al. (2013), where the winning teams using the test data achieved scores of 87.24, 84.61, and 83.19%, whereas the results reported in this article show classification accuracies from 86.4 to 92.6% using four different state-of-the-art classification algorithms. It is important to note that even when the CGD13 data set was used, the challenge itself was not part of the experiment. Each gesture used for testing had a label assigning

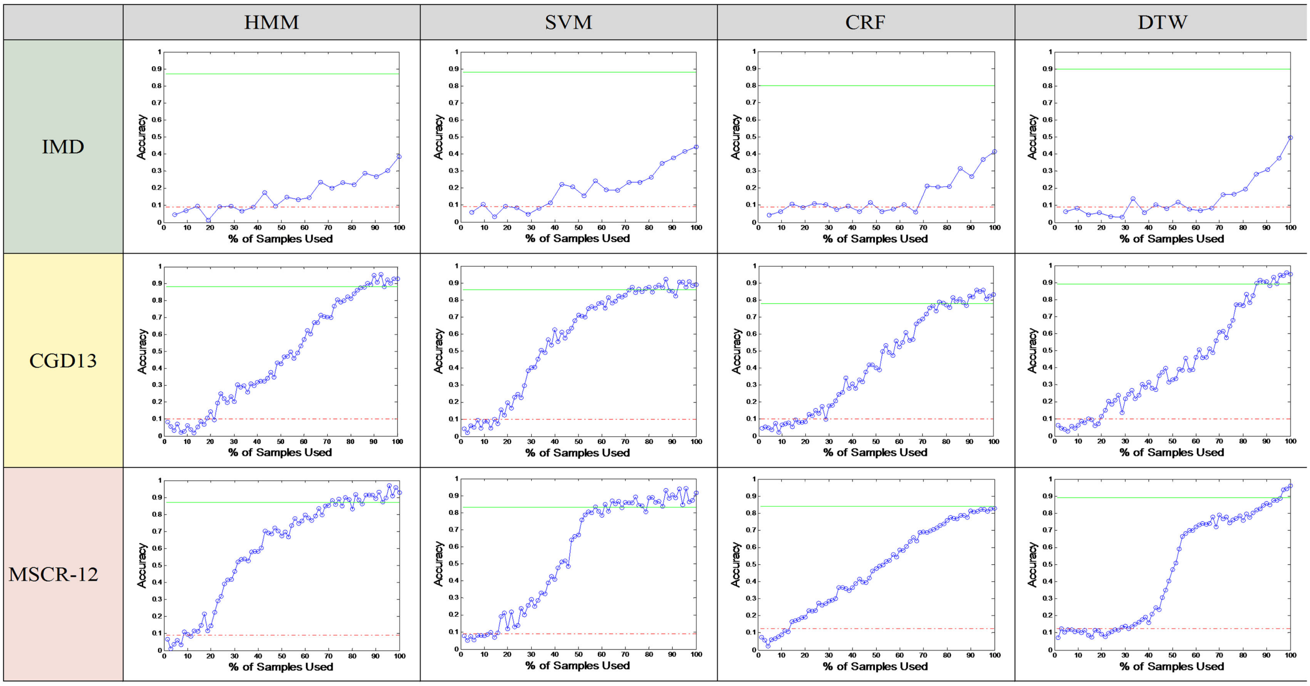


FIGURE 13 | Accuracy versus percentage of training samples for all combinations (data set–classifier).

TABLE 2 | Efficiency metric for two of the data sets.

η	Hidden Markov models	Support vector machines	Conditional random fields	Dynamic time warping
IMD	lim η → 1	lim η → 1	lim η → 1	lim η → 1
CGD13	0.982	0.979	0.981	0.983
Microsoft Research Cambridge-12	0.979	0.976	0.985	0.985

it to one of the 10 gesture classes selected for this particular experiment. No spotting technique was applied.

Regarding the MSCR-12 data set, previous results reported by Ellis et al. (2013) reached 88.7% accuracy, whereas Ramírez-Corona et al. (2013) achieved 91.82%. The proposed method achieved accuracies between 90.6% and 93.3%. Our experiment used only a subset of the data set and a subset of the gesture vocabulary. Regardless, the proposed one-shot learning approach used to train different classification algorithms gives results comparable to those of state-of-the art approaches that use multiple examples as training data.

Another interesting perspective related to our experiments is the validation of the approach when the gesture instances were performed by a dual-arm robotic platform. While the results were not as good as those obtained using skeleton information on human subjects, varying between 86.88% and 90% depending on the classification algorithm, this robotic implementation opens a different route toward coherency in human–machine interaction. This concept of coherency can be related to agreement metrics in gesture classification when the roles of performing and recognizing a gesture are interchanged between humans and machines.

As an efficiency metric, the recognition accuracy obtained previously was used as a baseline to determine the number of samples required to achieve similar recognition results to those from a traditional *N*-shot learning approach. This metric is related to the ability to save data acquisition time. This means reducing the time needed to acquire and process numerous training samples. This perspective on classification performance is an advance on current views of the one-shot learning problem. The generalizability obtained with the gist of a gesture approach is assumed to be similar to that obtained by increasing the number of samples required to match the recognition accuracy metric for each classifier.

In its application to one-shot learning, the proposed method highlights the use of context for gesture recognition from the way humans use their bodies. Future work will incorporate interjoint constraints in the kinematic chain of the human arms as a new method to use the gist of the gesture and thereby expand the number of samples from a gesture class to use as training data for classification algorithms.

ETHICS STATEMENT

This study was carried out with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the IRB (approved protocol #1609018129).

AUTHOR CONTRIBUTIONS

Both authors have participated in the process of designing the experiments and actively participated in the writing process of this manuscript.

FUNDING

This research is supported partly by the NPRP award (NPRP 6-449-2-181) from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors. This work is partly supported in part by the US Air Force Office of Scientific Research Young Investigator Research Program award no. FA 9550-13-1-0141 to JPW. This work is partly supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-14-1-0042. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

REFERENCES

Adams, J. A. (2005). "Human-robot interaction design: understanding user needs and requirements," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49 (SAGE Publications), 447–451. Available at: http://pro.sagepub.com/content/49/3/447.short

Albrecht, I., Haber, J., and Seidel, H.-P. (2003). "Construction and animation of anatomically based human hand models," in *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland: Eurographics Association), 98–109. Available at: http://dl.acm.org/citation.cfm?id=846276.846290

Andersen, D., Popescu, V., Cabrera, M. E., Shanghavi, A., Gomez, G., Marley, S., et al. (2016). Medical telementoring using an augmented reality transparent display. *Surgery* 159, 1646–1653. doi:10.1016/j.surg.2015.12.016

Babu, S. (2016). "A novel sparsity based classification framework to exploit clusters in data," in *Industrial Conference on Data Mining* (Springer International Publishing), 253–265.

Bellugi, U. (1979). *The Signs of Language*. Harvard University Press. Available at: https://books.google.com/books?hl=en&lr=&id=WeBOn6N8PJ8C&oi=fnd&pg=PA1&dq=The+Signs+of+Language,&ots=S97lt_MYL4&sig=JC03-e2iTVxgDXJ3iZvMe4su_f74

Bhuyan, M. K., Neog, D. R., and Kar, M. K. (2011). "Hand pose recognition using geometric features," in *2011 National Conference on Communications (NCC)* (Bangalore), 1–5.

Brown, A. L., and Kane, M. J. (1988). Preschool children can learn to transfer: learning to learn and learning from example. *Cogn. Psychol.* 20, 493–523. doi:10.1016/0010-0285(88)90014-X

Cabrera, M. E., Novak, K., Foti, D., Voyles, R., and Wachs, J. P. (2017a). "What makes a gesture a gesture? Neural signatures involved in gesture recognition," in *Accepted to 12th IEEE International Conference on Automatic Face and Gesture Recognition*. Available at: https://arxiv.org/abs/1701.05921

Cabrera, M. E., Voyles, R., and Wachs, J. P. (2017b). *Coherency in Gesture Recognition*. Available at: https://arxiv.org/abs/1701.05924

Cabrera, M. E., and Wachs, J. P. (2016). "Embodied gesture learning from one-shot," in *The 25th IEEE International Symposium on Robot and Human Interactive Communication* (New York: IEEE), 1092–1097.

Caulfield, H. J., and Heidary, K. (2005). Exploring margin setting for good generalization in multiple class discrimination. *Pattern Recognit.* 38, 1225–1238. doi:10.1016/j.patcog.2005.01.009

ChaLearn Looking at People. (2014). *Accessed June 4*. Available at: http://gesture.chalearn.org/

CRF++: Yet Another CRF Toolkit. (2016). *Accessed July 1*. Available at: https://taku910.github.io/crfpp/

Ellis, C., Masood, S. Z., Tappen, M. F., Laviola, J. J. Jr, and Sukthankar, R. (2013). Exploring the trade-off between accuracy and observational latency in action recognition. *Int. J. Comput. Vis.* 101, 420–436. doi:10.1007/s11263-012-0550-7

Escalante, H. J., Guyon, I., Athitsos, V., Jangyodsuk, P., and Wan, J. (2017). Principal motion components for one-shot gesture recognition. *Pattern Anal. Appl.* 20, 167–182. doi:10.1007/s10044-015-0481-3

Escalera, S., Athitsos, V., and Guyon, I. (2016). Challenges in multimodal gesture recognition. *J. Mach. Learn. Res.* 17, 1–54.

Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., et al. (2013). "Multi-modal gesture recognition challenge 2013: dataset and results," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (ACM), 445–452. Available at: http://dl.acm.org/citation.cfm?id=2532595

Fanello, S. R., Gori, I., Metta, G., and Odone, F. (2013). Keep it simple and sparse: real-time action recognition. *J. Mach. Learn. Res.* 14, 2617–2640.

Fe-Fei, L., Fergus, R., and Perona, P. (2003). "A bayesian approach to unsupervised one-shot learning of object categories," in *Proceedings of the Ninth IEEE International Conference on Computer Vision* (IEEE), 1134–1141. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1238476

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 594–611. doi:10.1109/TPAMI.2006.79

Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. (2012). "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM), 1737–1746. Available at: http://dl.acm.org/citation.cfm?id=2208303

Gesture Recognition Toolkit. (2016). Available at: http://www.nickgillian.com/wiki/pmwiki.php/GRT/GestureRecognitionToolkit

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends Cogn. Sci.* 3, 419–429. doi:10.1016/S1364-6613(99)01397-2

Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H. J., and Hamner, B. (2013). "Results and analysis of the chalearn gesture challenge 2012," in *Advances in Depth Image Analysis and Applications* (Springer), 186–204. Available at: http://link.springer.com/chapter/10.1007/978-3-642-40303-3_19

Guyon, I., Athitsos, V., Jangyodsuk, P., Hamner, B., and Escalante, H. J. (2012). "Chalearn gesture challenge: design and first results," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference* (IEEE), 1–6. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6239178

Hertz, T., Hillel, A. B., and Weinshall, D. (2006). "Learning a kernel function for classification with small training samples," in *Proceedings of the 23rd International Conference on Machine Learning* (New York, NY: ACM), 401–408.

Hewes, G. W. (1992). Primate communication and the gestural origin of language. *Curr. Anthropol.* 33, 65–84. doi:10.1086/204019

Huang, C., Ai, H., Li, Y., and Lao, S. (2007). High-performance rotation invariant multiview face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 671–686. doi:10.1109/TPAMI.2007.1011

Ikizler, N., and Duygulu, P. (2009). Histogram of oriented rectangles: a new pose descriptor for human action recognition. *Image Vis. Comput.* 27, 1515–1526. doi:10.1016/j.imavis.2009.02.002

Jacob, M. G., and Wachs, J. P. (2014). Context-based hand gesture recognition for the operating room. *Pattern Recognit. Lett.* 36, 196–203. doi:10.1016/j.patrec.2013.05.024

Jaimes, A., and Sebe, N. (2007). Multimodal human–computer interaction: a survey. *Comput. Vis. Image Underst.* 108, 116–134. doi:10.1016/j.cviu.2006.10.019

Jiang, H., Huang, K., Mu, T., Zhang, R., Ting, T. O., and Wang, C. (2016a). Robust one-shot facial expression recognition with sunglasses. *Int. J. Mach. Learn. Comput.* 6, 80–86. doi:10.18178/ijmlc.2016.6.2.577

Jiang, H., Duerstock, B. S., and Wachs, J. P. (2016b). User-centered and analytic-based approaches to generate usable gestures for individuals with quadriplegia. *IEEE Trans. Hum. Mach. Syst.* 46, 460–466. doi:10.1109/THMS.2015.2497346

Johnson, J. E., Christie, J. F., and Wardle, F. (2005). *Play, Development, and Early Education*. Pearson/Allyn and Bacon. Available at: https://ghnet.guelphhumber.ca/files/course_outlines/ECS2020_outline_W12_Pascarielle.approved.pdf

Jost, C., De Loor, P., Nédélec, L., Bevacqua, E., and Stanković, I. (2015). "Real-time gesture recognition based on motion quality analysis," in *2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)* (Torino), 47–56.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. *Relat. Verbal Nonverbal Commun.* 25, 207–227.

Kendon, A. (1986). Current issues in the study of gesture. *Biol. Found. Gestures Mot. Semiotic Aspects* 1, 23–47.

Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*. 7.

Kendon, A. (1994). Do gestures communicate? A review. *Res. Lang. Soc. Interact.* 27, 175–200. doi:10.1207/s15327973rlsi2703_2

Konecny, J., and Hagara, M. (2014). One-shot-learning gesture recognition using hog-hof features. *J. Mach. Learn. Res.* 15, 2513–2532.

Kopicki, M., Detry, R., Adjigble, M., Stolkin, R., Leonardis, A., and Wyatt, J. L. (2016). One-shot learning and generation of dexterous grasps for novel objects. *Int. J. Robot. Res.* 35, 959–976. doi:10.1177/0278364915594244

Krauss, R. M., Morrel-Samuels, P., and Colasante, C. (1991). Do conversational hand gestures communicate? *J. Pers. Soc. Psychol.* 61, 743. doi:10.1037/0022-3514.61.5.743

Kwitt, R., Hegenbart, S., and Niethammer, M. (2016). *One-Shot Learning of Scene Locations via Feature Trajectory Transfer*. 78–86. Available at: http://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Kwitt_One-Shot_Learning_of_CVPR_2016_paper.html

Lake, B. M., Salakhutdinov, R., Gross, J., and Tenenbaum, J. B. (2011). "One shot learning of simple visual concepts," in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (Austin), 2568–2573.

Lee, C.-S., Elgammal, A., and Torki, M. (2016). Learning representations from multiple manifolds. *Pattern Recognit.* 50, 74–87. doi:10.1016/j.patcog.2015.08.024

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Phys. Dokl.* 10, 707.

Liang, R.-H., and Ouhyoung, M. (1998). "A real-time continuous gesture recognition system for sign language," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition* (Nara), 558–567.

Maas, A., and Kemp, C. (2009). *One-Shot Learning with Bayesian Networks*. Cognitive Science Society. Available at: http://repository.cmu.edu/psychology/972/

Mapari, R. B., and Kharat, G. (2015). "Real time human pose recognition using leap motion sensor," in *2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* (Kolkata), 323–328.

Marron, J. S., Todd, M. J., and Ahn, J. (2007). Distance-weighted discrimination. *J. Am. Stat. Assoc.* 102, 1267–1271. doi:10.1198/016214507000001120

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press. Available at: https://books.google.com/books?hl=en&lr=&id=3ZZAfNumLvwC&oi=fnd&pg=PA6&dq=hand+and+mind&ots=oId6REwEaz&sig=DoUn_unvf0OeFdp3detGKqMNhXQ

McNeill, D., and Levy, E. (1980). *Conceptual Representations in Language Activity and Gesture*. Available at: http://eric.ed.gov/?id=ED201202

Merler, M., Huang, B., Xie, L., Hua, G., and Natsev, A. (2012). Semantic model vectors for complex video event recognition. *IEEE Trans. Multimedia* 14, 88–101. doi:10.1109/TMM.2011.2168948

Nickel, K., and Stiefelhagen, R. (2007). Visual recognition of pointing gestures for human–robot interaction. *Image Vis. Comput.* 25, 1875–1884. doi:10.1016/j.imavis.2005.12.020

Ormrod, J. E., and Davis, K. M. (2004). *Human Learning*. Merrill. Available at: http://www.nhmnc.info/wp-content/uploads/fbpdfs2014/Human-Learning-6th-Edition-by-Jeanne-Ellis-Ormrod-If-The-Book-Was-Good-Wouldnt-The-Publisher-Let-You-Search-It-.pdf

Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). "Zero-shot learning with semantic output codes," in *Advances in Neural Information Processing Systems 22*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Curran Associates, Inc.), 1410–1418. Available at: http://papers.nips.cc/paper/3650-zero-shot-learning-with-semantic-output-codes.pdf

Parikh, D., and Grauman, K. (2011). "Interactively building a discriminative vocabulary of nameable attributes," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE), 1681–1688. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5995451

Pavlovic, V. I., Sharma, R., and Huang, T. S. (1997). Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 677–695. doi:10.1109/34.598226

Ramírez-Corona, M., Osorio-Ramos, M., and Morales, E. F. (2013). "A non-temporal approach for gesture recognition using microsoft kinect," in *Iberoamerican Congress on Pattern Recognition* (Springer), 318–325. Available at: http://link.springer.com/chapter/10.1007/978-3-642-41827-3_40

Rautaray, S. S., and Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artif. Intell. Rev.* 43, 1–54. doi:10.1007/s10462-012-9356-9

Rekabdar, B., Nicolescu, M., Nicolescu, M., and Kelley, R. (2015). "Scale and translation invariant learning of spatio-temporal patterns using longest common subsequences and spiking neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)* (Anchorage), 1–7.

Rigoll, G., Kosmala, A., and Eickeler, S. (1997). "High performance real-time gesture recognition using hidden markov models," in *Gesture and Sign Language in Human-Computer Interaction*, eds W. Ipke and M. Fröhlich (Berlin, Heidelberg: Springer), 69–80. Available at: http://link.springer.com/chapter/10.1007/BFb0052990

ROS.org | Powering the World's Robots. (2016). Available at: http://www.ros.org/

Sarup, P., Jensen, J., Ostersen, T., Henryon, M., and Sørensen, P. (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred danish duroc pigs. *BMC Genet.* 17:11. doi:10.1186/s12863-015-0322-9

Serre, T., Wolf, L., and Poggio, T. (2005). "Object recognition with features inspired by visual cortex," in*IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2 (IEEE), 994–1000. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1467551

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. (2013). "Zero-shot learning through cross-modal transfer," in *Advances in Neural Information Processing Systems 26*, eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc.), 935–943. Available at: http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf

Sun, Y., Liang, D., Wang, X., and Tang, X. (2015). *DeepID3: Face Recognition with Very Deep Neural Networks." arXiv:1502.00873 [Cs]*. Available at: http://arxiv.org/abs/1502.00873

Toshev, A., Makadia, A., and Daniilidis, K. (2009). "Shape-based object recognition in videos using 3D synthetic object models," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), 288–295. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5206803

Wan, J., Ruan, Q., Li, W., and Deng, S. (2013). One-shot learning gesture recognition from RGB-D data using bag of features. *J. Mach. Learn. Res.* 14, 2549–2582.

Wang, D., Nie, F., and Huang, H. (2015). Feature selection via global redundancy minimization. *IEEE Trans. Knowl. Data Eng.* 27, 2743–2755. doi:10.1109/TKDE.2015.2426703

Wasikowski, M., and Chen, X. W. (2010). Combating the small sample class imbalance problem using feature selection. *IEEE Trans. Knowl. Data Eng.* 22, 1388–1400. doi:10.1109/TKDE.2009.187

Wei, J., Jian-qi, Z., and Xiang, Z. (2011). Face recognition method based on support vector machine and particle swarm optimization. *Expert Syst. Appl.* 38, 4390–4393. doi:10.1016/j.eswa.2010.09.108

Wu, D., Zhu, F., and Shao, L. (2012). "One shot learning gesture recognition from RGBD images," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Providence), 7–12.

Yamato, J., Ohya, J., and Ishii, K. (1992). "Recognizing human action in time-sequential images using hidden markov model," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE), 379–385. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=223161

Yang, H.-D., Park, A.-Y., and Lee, S.-W. (2007). Gesture spotting and recognition for human–robot interaction. *IEEE Trans. Robot.* 23, 256–270. doi:10.1109/TRO.2006.889491

Zaffalon, M., and Hutter, M. (2002). "Robust feature selection by mutual information distributions," in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 577–584. Available at: http://dl.acm.org/citation.cfm?id=2073876.2073945

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *Multimedia IEEE* 19, 4–10. doi:10.1109/MMUL.2012.24

Zheng, J., Wang, Y., and Zeng, W. (2015). "CNN based vehicle counting with virtual coil in traffic surveillance video," in *2015 IEEE International Conference on Multimedia Big Data (BigMM)* (Beijing), 280–281.

Zhou, E., Cao, Z., and Yin, Q. (2015). *Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?" arXiv:1501.04690 [Cs]*. Available at: http://arxiv.org/abs/1501.04690

Conflict of Interest Statement: 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.