frontiers in Computer Science

ORIGINAL RESEARCH
published: 10 January 2022
doi: 10.3389/fcomp.2021.787735

Communicating Photograph Content Through Tactile Images to People With Visual Impairments

Karolina Pakėnaitė[1]*, *Petar Nedelev*[1], *Eirini Kamperou*[2], *Michael J. Proulx*[2] *and Peter M. Hall*[1]

[1]*Department of Computer Science, University of Bath, Bath, United Kingdom,* [2]*Department of Psychology, University of Bath, Bath, United Kingdom*

Millions of people with a visual impairment across the world are denied access to visual images. They are unable to enjoy the simple pleasures of viewing family photographs, those in textbooks or tourist brochures and the pictorial embellishment of news stories etc. We propose a simple, inexpensive but effective approach, to make content accessible *via* touch. We use state-of-the-art algorithms to automatically process an input photograph into a collage of icons, that depict the most important semantic aspects of a scene. This collage is then printed onto swell paper. Our experiments show that people can recognise content with an accuracy exceeding 70% and create plausible narratives to explain it. This means that people can understand image content *via* touch. Communicating scene foreground is a step forward, but there are many other steps needed to provide the visually impaired with the fullest possible access to visual content.

Keywords: communicating photograph content, visual impairments, accessibility, tactile image recognition, icons, tactile representations, automatic object detection, empirical verification

OPEN ACCESS

Edited by:
Antonino Furnari,
University of Catania, Italy

Reviewed by:
Tyler Highlander,
Wittenberg University, United States
Farhan Mohamed,
University of Technology Malaysia,
Malaysia

Correspondence:
Karolina Pakėnaitė
kp421@bath.ac.uk

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

Received: *01 October 2021*
Accepted: *19 November 2021*
Published: *10 January 2022*

Citation:
Pakėnaitė K, Nedelev P, Kamperou E, Proulx MJ and Hall PM (2022) Communicating Photograph Content Through Tactile Images to People With Visual Impairments.
Front. Comput. Sci. 3:787735.
doi: 10.3389/fcomp.2021.787735

1 INTRODUCTION

Pictures are an important means of communication. Illustrations in newspapers or in books help bring a story to life. Photographs are used socially too, with around two billion being uploaded onto social media platforms every day[1]. Pictures can also be informative, with school textbooks, holiday brochures, museum catalogues, and more being dependent on them.

Our aim is to make photographic content accessible to people with a visual impairment at a low cost. This can build bridges of opportunity and interaction for people with a visual impairment, both professionally and socially (Way and Barner, 1997). Tactile images are currently made manually with the help of a skilled, sighted individual, which is neither cost effective nor time efficient[2].

Combining existing neural algorithms was explored. We used our developed algorithm to automatically process an input photograph into a collage of recognisable icons (similar to road signs). These icons can be simply collected from public online sources and are easy to read as classification symbols. The collage of black icons on white background is printed on swell paper to

[1] Wu, S., Pique, H., and Wieland, J. (2016) About Facebook. *Using Artificial Intelligence to Help Blind People "See" Facebook.* Retrieved from https://about.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook.
[2] Living Paintings. *Support us - Living Paintings.* Retrieved from https://livingpaintings.org/support-us.

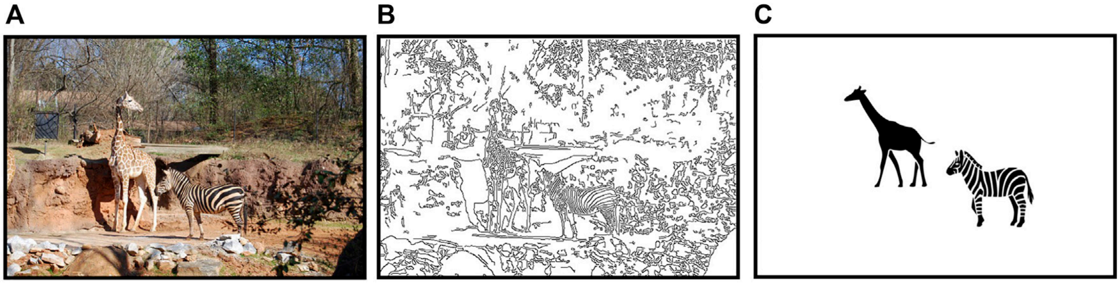

FIGURE 1 | Which photographic representation is easier to read into? The figure shows a panel of images with a photograph of animals in the **(A)**, an edge map of the same image in the **(B)**, and icons representing only the two animals in the **(C)**.

form a tactile rendering of the photograph. It is a tactile artwork that represents the most salient, identifiable objects in the picture.

To test the efficacy of our algorithm, experiments were conducted to compare the collage with an edge map representation. Canny edge detector was used, which is an Image Processing Algorithm that creates white pixels at window centres, if the window pattern contains a sufficiently strong contrast edge (Canny, 1986). Edge maps are easy to make and apply to any image, which makes them very cheap. However, they are difficult (but not impossible) for sighted people to read content into; the same is expected for tactile renderings of edge maps. For the same photograph input in **Figure 1A**, the difference can be seen between an edge-map representation **Figure 1B** and our icon-based representation **Figure 1C**. **Figure 1** illustrates the difference between our icon-based representation and an edge map representation for the same input photograph.

Research participants (32 blindfolded, consisting of 12 with visual impairments) were asked to describe the content of 10 tactile images taken from photographs of 12 different objects. Results show that icons convey more information about input photographs than edge maps; people can describe content with a far greater degree of accuracy. Moreover, people were more likely to comment on, or compose stories about icon-images than edge map images.

We have shown that icon based tactile images are successful in conveying semantic content, but more work is needed to fully realise the potential of our general approach. Background context is missing, which intuitively is a low barrier to break. Many improved alternatives to edge maps also exist, such as boundary of a segmented object. The optimal way to convey information about a photo in tactile form remains an open question.

2 BACKGROUND

There are 285 million people worldwide with a visual impairment (Pascolini and Mariotti, 2011) and many are regularly faced with the challenge of accessing visual information. This includes activities such as viewing and sharing photographs online, which has become a modern way of social interaction. Textbook content might be quicker and easier to understand by the sighted after viewing the corresponding image. Offline or online, visually impaired users wish to view such images independently.

Consider **Figure 1A** above. How might a viewer describe this image to someone who cannot see it? Which aspects are crucial for conveying the information, and what is the appropriate amount of description? Different approaches exist to help solve this problem. One might give a simple verbal description, or a detailed interpretation. With advances in technology, many have adapted to a screen reader-an automated software that verbally describes images using speech. However, this type of software might not be helpful for everyone, especially individuals with an additional disability such as a hearing impairment. Alternatively, some have adjusted to accessing visual information by feeling tactile images with their fingers.

Artists' work generally includes semantic details only. This observation motivated us to develop an algorithm for rendering tactile pictures by simplifying both the low-level features and the higher-level semantic content of the input. This was done both visually and logically, whilst preserving the original context. The aim is to increase tactile comprehension, and reconstruct an original context of a photograph for a visually impaired user, by simplifying the input image and outputting the most important semantic contents.

To ensure consistency of terminology and to avoid confusion, two terms will be used to describe people with a visual impairment based on the World Health Organization definition of blindness. By looking at visual acuity, if an individual has a reduced visual input (logMAR < 1.3, $n = 15$), the term Low Vision is used. If an individual has little or no functional visual input (logMAR ≥ 1.3; $n = 14$), the term Blind is used[3]. The term Visually Impaired will be used for both groups - those who are blind and low vision. 8 years was chosen as a cut-off age to differentiate between early and late onset, with congenitally blind being included in the early group. The cut-off age was based

[3] World Health Organization. (2021). International statistical classification of diseases and related health problems (11th Revision). *9D90 Vision impairment including blindness*. Retrieved from https://icd.who.int/browse11/l-m/en.

on a study where primary visual cortex was mostly active among the late blind (Büchel et al., 1998).

2.1 Psychology of Tactile Image Recognition

A tactile picture is a physical representation of an image, executed in relief and usually printed on a special paper, with swollen or raised parts that can be felt with the fingers. Swell paper has a layer of heat reactive microcapsules, and when a heat fuser is used, the black markings are raised. Printing on swell paper can be a cost-effective option and is suitable for use with a standard printer. There are several techniques with different features for printing tactile pictures, for example some allow single, and others multiple levels of relief (Eriksson, 1999). The specific technique for producing the image is irrelevant as only the general case will be tackled here.

Tactile representations which allow visually impaired individuals to perceive images in relief, needs some guarantee for the content to be understood. Eriksson (1999) outlines the form of a tactile picture, so that semantics can be perceived easily by touch. The most important characteristic for identifying an object is usually the shape, (not the colour or material) so all objects within a tactile picture require separating without overlap. The components of an object need to be in a distinct and logically simplified form, with no incomplete objects.

When looking at photographs, it is generally the context that is accessed first (Oliva and Torralba, 2006). This is usually true for blind individuals when accessing information in photographs online, or art pieces in museums, using touch or verbal descriptions (Stangl et al., 2020; Hayhoe, 2013). Eriksson (1999) further elaborates that it is rare for a visually impaired individual to comprehend content or identify objects when touching a tactile picture for the first time. This is due to the unique shape of each object being difficult to classify without prior experience of the shape. This means that most tactile pictures should be accompanied with a description, to facilitate the reader's understanding of what they are feeling. This is not required if the reader has had previous exposure to the shapes and objects.

Objects represented in 3D on 2D media, can be a heavy burden on memory (Biederman, 1987). One study revealed that sighted individuals recognise 3D drawings slower and with less accuracy than 2D representations, because additional processing is required (Lederman et al., 1990). In addition, tactual performance of congenitally blind study participants was identical in both two and three-dimensional representations on 2D media. This is because interpretation of the third dimension is difficult without visual imagery. Thus, icons represented in 2D were chosen for the rendering stage. However, it is also observed that sighted, blindfolded participants only achieved around 33% success when attempting to recognise common objects that were represented as raised 2D outlines (Lederman et al., 1990). In contrast, 20 participants successfully recognised 100 common objects by touch within 1 or 2 s. It concluded that it is important to incorporate diagnostic substance differences on raised pictures (Klatzky et al., 1985).

A novel technique for producing 3D objects in 2D media as tactile line illustrations, has obtained promising results in object-matching by blind participants (Panotopoulou et al., 2020). Object models were taken at different, carefully selected angles to extract 3D object parts with added texturing technique, to give cues about the surface geometry. The illustration of 3D objects was then printed onto microcapsule paper. Multi-projection approach was inspired by drawings created by blind individuals (Kennedy, 1993), where 3D objects are "unfolded" or "flattened" to obtain 2D illustrations (Kurze, 1997). However, it was not confirmed if this improved object recognition without matching.

Each detected object in an image should be represented solely by a single, fixed-pose icon. This was partly supported by a study where participants showed a preference for three-quarters front view, when recognising objects (Palmer, 1981). Bartram (1974) added that sighted individuals are slower to recognise objects of different forms and shapes (e.g., a table lamp and a floor lamp), than when viewing identical objects in different views (e.g., viewing a table lamp from the left or right side). Results from another study showed that congenitally blind subjects can take the point of view of another individual, and that visual experience is not necessary (Heller and Kennedy, 1990).

Sighted individuals tend to spend longer exploring tactile images, and therefore experience an excessive burden on memory. In addition, blind individuals who understand braille, tend to use different exploration strategies (i.e., using two fingers instead of one), which can result in a faster pick up of information and less memory burden (Heller, 1989). One study confirmed that in raised line pictures, identification improved when five fingers were used instead of one. It also added that multiple fingers could potentially have a role in guiding exploration, when the field of touch-view is increased (Klatzky et al., 1993). A few reports state that people lose touch sensitivity as they age (Manning and Tremblay, 2006; Tremblay and Master, 2015). However, continuous practice in tactile recognition seems to compensate for deteriorating sensitivity associated with advancing age (Legge et al., 2008). It could be assumed that tactile images are better comprehended by braille users than sighted individuals, due to their experience of touching textured materials. Another study concluded that individuals who lose vision later in life, tend to interpret raised outlines more readily than sighted individuals, or those who became blind before the age of 3 months. An explanation being that they are typically more familiar with pictures than early onset blind individuals, and have greater tactile skills than the sighted. It also highlighted that visual experience alone does not increase the performance of a tactile task, since sighted participants did not score better than the congenitally blind (Heller, 1989).

As mentioned earlier, there are important issues related to dealing with information processing and potential cognitive or perceptual load (Gallace and Spence, 2009; Lin et al., 2021). The success of delivering visual information through touch, is not only dependent on the technical specifics of the sensory substitution device, but also on understanding how the brain processes multi-sensory information (Brown and Proulx, 2016).

2.2 Computer Vision Techniques for Making Tactile Images

Tactile images can be made by hand. These are high quality and often use raised contour lines or bas-relief regions, possibly with colour projected onto the surface. This kind of tactile image often exists for access to Fine Art and other cultural artefacts (Reichinger et al., 2018; Cantoni et al., 2018). However, hand-made, bespoke tactile images can be expensive one-offs, so there is a need for an inexpensive and more general approach.

One automatic method for creating a tactile image is to use edge maps, see **Figure 1B** for an example. This kind of tactile image is exceptionally easy to create. The Canny edge detector (Canny, 1986) for example, takes a computer fractions of a second to produce an edge map. However, edge maps are not an efficient way to communicate in tactile form (Way and Barner (1997), Eriksson (1999)) and are only acceptable for the simplest of objects.

We use Computer Vision algorithms to automatically process the important semantic content of an input photograph, into a form that can be printed as tactile image. There are many state-of-the-art algorithms to choose from based around neural networks.

Many networks are able to detect objects in photographs, including but not limited to Romera-Paredes and Torr (2016); Salvador et al. (2017); Carion et al. (2020). Each network solves the problem in slightly different ways. The variations are of importance to Computer Vision researchers, but for our purposes we have the luxury of choice. We opt for Mask R-CNN (He et al., 2017) because it: 1) makes code available to be used; 2) is fast to run; 3) reliably detects many kinds of things; 4) works on ordinary photographs (consisting multiple objects in a complex environment) as well as on photographs in a data set (often showing one object only); 5) detects objects at pixel level, outputting silhouette rather than by placing a box containing an object on the image.

Object segmentation is not sufficient for us; we need to know which objects are important. Some object detectors claim to segment salience objects, for example see work by Romera-Paredes and Torr (2016), but we opt to compute salience independently. The pixel-level detector we use means we can combine it with a pixel-level salience detector.

Saliency Detection can be defined as predicting the human fixation points in images, or as identifying distinct visual regions or objects within an image (Borji et al., 2015). The latter is entirely subject to Bottom-up Saliency, whereas human fixation points can be influenced both by Bottom-up Saliency and by Top-down Control (defined by the semantic relationships between objects within an image, and the current inner goal) (Melloni et al., 2012). However, eye tracking techniques (Rossi et al., 2017) would be neither practical, nor feasible for this project. We have opted to use PiCANet (Liu et al., 2020), where for each image, pixel informative contextual regions are learnt and contextual attention is generated, thus allowing useful semantic features to be obtained and better decisions made.

2.3 Non-Visual Methods

Before leaving the background section The use of non tactile methods that communicate visual images in a non visual form are mentioned which communicates visual images in a non-visual form.

Many people with a visual impairment have adapted to technology and use Alt Text when navigating the internet through a screen reader. Alt Text produces short descriptions of photographs in written form. Facebook Alt Text (Wu et al., 2017) is an example of this and is used today. It works automatically, saving time and human effort. Currently, Facebook Alt Text lists basic information and descriptions in a particular order as follows: people (e.g., whether it is a person smiling, a baby or a child), objects (e.g., car, cloud, dog, shoes), and settings/themes (e.g., inside restaurant, outdoors, in nature). Facebook Alt Text recognises 100 common objects, and is not written in sentences to keep accuracy as high as possible.

Just as sighted people benefit from a combination of prose and pictures, visually impaired people might too. Our provision of tactile images complements, rather than competes with alt-text and other non-visual, non-tactile methods.

3 OUR ALGORITHM: FROM VISUAL TO TACTILE IMAGE

One of the contributions of this paper is developing an algorithm to process photographs into a tactile rendering. Our form of tactile rendering is a swell-paper image of a collage containing black icons on white background. The algorithm outputs a collage of icons that reflects the salient content of an input photograph.

As an overview, our algorithm has two parts, as shown in **Figure 2**. First, the input photo is analysed into a map of the most salient objects in the picture. The map locates the objects at pixel level, so each can appear in silhouette. Each object has been recognised as an object and each has a unique and a number to identify it. See **Section 3.1** for details.

This map is used as the starting point for the second section, which is to create a collage of icons. Each identified salient object is replaced with its corresponding icon, which is scaled to size and moved if necessary to avoid overlap with nearby icons. The details of this are left to **Section 3.2**.

3.1 Analysis

The purpose of analysis is to construct a model of the input photograph, which is a map of salient objects. We use two neural networks for this. The first, Mask R-CNN (He et al., 2017), detects objects in the photograph. The output is a map of object labels, $object(x, y)$. When object detection is running, weights are loaded from a model trained on the MS-COCO dataset (Lin et al., 2014) into the network. The dataset consists of images representing 81 different classes of common objects (e.g., a person or a bike), meaning that this is the number of classes the network can recognise and classify.

The second network, PiCANet (Liu et al., 2020), provides a salience map. We use a bottom-up salience map, typically obtained from eye-gaze experiments: the algorithm assumes that dwell time correlates with salience, and it learns to

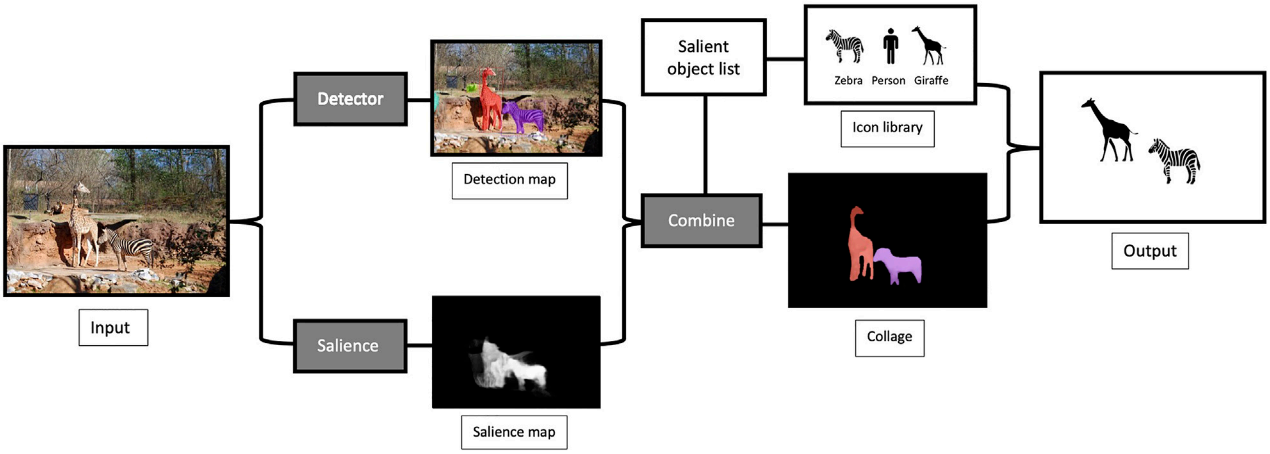

FIGURE 2 | The algorithm in overview, from visual to tactile images. The input image at the left is processed to produce both a detection and a salience map. These are combined into a collage, with a salient object list to call up the icons in the library. The output image places the relevant icons into representative spatial locations in the image.

associate spatial colour patterns with dwell time. The output is a salience map, *sal*(*x*, *y*), that is brightest over the most salient pixels.

We now have a combined map of the form:

$$map(x, y) = [object(x, y), sal(x, y)]. \quad (1)$$

We also have a list of the objects in the picture, which provides the information to make a collage.

3.2 Rendering

The first step in rendering is to decide which objects to keep, i.e., which are the most salient. This is straightforward, salience per unit area for the j^{th} object is computed as:

$$s(j) = \frac{\sum_{xy} sal(x, y)(id(x, y) = j)}{\sum_{xy}(id(x, y) = j)}. \quad (2)$$

Next, the objects are thresholded to retain only the most salient. Choosing a threshold is not easy; the intention is to find balance between too many and too few objects. Simple thresholds, such as using the mean of all salient objects does not succeed this balance. A heuristic solution is using the mean μ, and deviation σ of all salience values. We therefore define the threshold as:

$$\tau(\mu, \sigma) = \begin{cases} \mu, & \sigma > \mu \\ \mu - 2\sigma, & \sigma < 10 \\ \mu - \frac{\sigma}{2}, & \text{otherwise.} \end{cases} \quad (3)$$

Each detected object has a corresponding icon within an assembled library. All icons are in png format, have equal dimensions and are used to create the collage. Any selected icon must be scaled to the correct size and moved, so it sits as closely as possible over the corresponding object, given the other objects around it.

That is, the icons are scaled by the ratio of the bounding rectangle perimeter of the segmented object to the perimeter of the icon: $s = perimeter(object)/perimeter(icon)$. The scaling is uniform so the shape of the icon does not change.

If the bounding rectangles of two objects overlap, then two conditions arise. In the first case, the rectangle of one object sits entirely inside the rectangle of another. Therefore, the smaller object is deleted from further consideration. Had this not been done, the icons could combine to make a single shape that would be difficult to recognise. In the second case, the boxes partially overlap. Here, the least salient of the two objects is pushed in the direction of a vector that joins the box centres until there is no overlap. This preserves the location of the most salient object, and ensures that the icons cannot overlap.

Moving a box away from another might lead to one box leaving the constraints of the image, thus a scaling operation is applied. If a given bounding box reaches the border of the image, the box is simply resized until it no longer overlaps any other boxes. In an image, there could be many overlapping boxes. In an attempt to retain as much of the original position of the objects as possible, of the original position of the objects, we iterate through the sets of all pairs of overlapping boxes, and at each iteration only a single move per pair is allowed.

The output image is then rendered as black icons on white background and printed on swell paper for later use.

4 EXPERIMENTS: ICONS ARE EFFECTIVE TACTILE REPRESENTATIONS

A mixed repeated measures design was used to assess the number of correctly identified objects in each image, with familiar and unfamiliar objects for edge maps and icons only representations.

4.1 Participants

A statistical power analysis was performed for sample size estimation, using data from a published study by Lederman

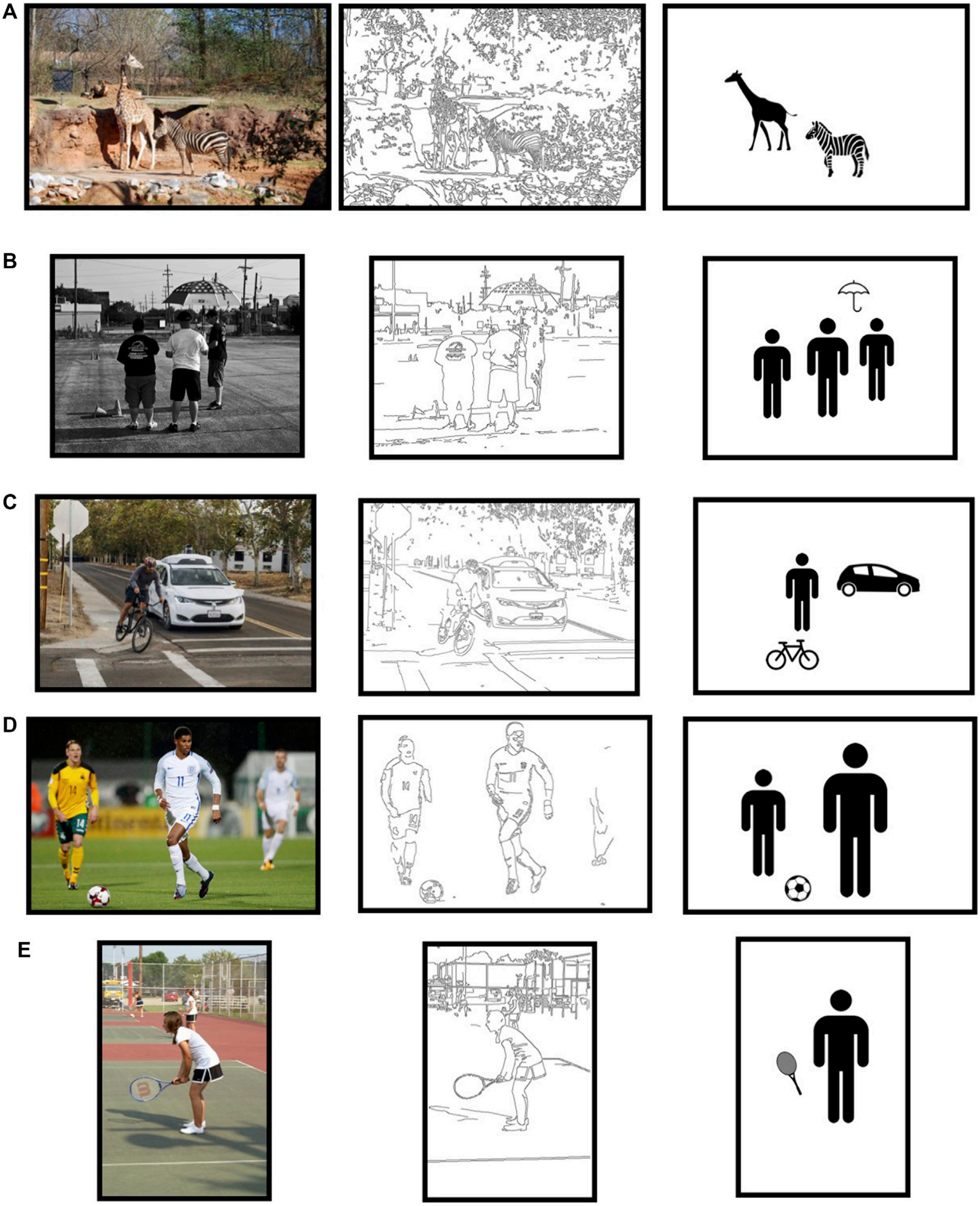

FIGURE 3 | Examples of photographs being represented as an edge map or using icons only.

et al. (1990). Performance differences for tactually identifying two and three-dimensional drawings of common objects were compared between seven sighted blindfolded, and seven congenitally blind participants. The effect size in this study was 1.5, which is considered extremely large using the criteria by Cohen (2013). With $\alpha = 0.05$ and *power* = 0.8, the projected sample size needed with this effect size [when using G*Power 3.1 (Faul et al., 2009)] is approximately $N = 14$ for this simplest between/within group comparison. Thus, our proposed sample size of $N + 6$ should be more than adequate for the main objective of this study. This should also allow for expected attrition, and our additional objectives of controlling for possible mediating/moderating factors/subgroup analysis, etc. Because it is more difficult to find a sufficient number of participants with visual impairment than sighted participants, the value of 0.2 was inputted into "Allocation ratio N2/N1" on G*Power 3.1. The output resulted in 18 for sample size group 1 and 4 for sample size group 2. Thus, the aim was to have at least 20 sighted and 10 visually impaired participants.

A total of 32 participants were recruited for the study, 12 of whom were visually impaired (37.5%). They were recruited by word of mouth from the University of Bath, the Association of the Blind of Western Greece, and Galloway's Society for the Blind. The study included 22 females (68.7%) and 10 males (31.3%), ranging from 16 to 62 years old ($M = 35.78$, $SD = 15.52$). Of the visually impaired participants, 2 were congenitally blind, 4 became blind in later life, and 6 were early blind. The average age was 41.33 years ($SD = 13.14$), while for the sighted participants the average age was 32.45 ($SD = 16.18$). Early blind participants had very little to no visual memory. 22 participants (68.8%) were not at all proficient in braille or moon system, five participants (15.6%) were not very proficient, 1 (3.1%) was quite proficient, 2 (6.3%) were proficient, and 2 (6.3%) were very proficient. Regarding tactile diagram experience, 22 participants (75.0%) were not at all proficient, 2 (6.3%) were not very proficient, 4 (12.5%) were quite proficient, 1 (3.1%) was proficient, and another 1 (3.1%) was very proficient.

4.2 Stimuli

The stimuli were created based on 10 photographs containing common objects. For icon only representations, a combination of object Mask R-CNN (He et al., 2017) and saliency PiCANet (Liu et al., 2020) detection algorithms were used to identify the most salient objects in each photograph and replaced to scale with an icon on a blank canvas. An example can be seen in the third column of **Figure 3**. When Mask R-CNN was running, weights were loaded from a model trained on the MS-COCO dataset into the network to detect any of 81 different classes of common objects (Lin et al., 2014). Object Detection architecture was implemented using Detectron (Facebook AI Research's software system)[4]. Saliency Detection PiCANet was implemented using MATLAB with the Caffe framework (Liu et al., 2020). An overall set of 12 icons were used, resulting from the detection of 22 salient objects. Eight icons of football, car, cat, dog, elephant, person, tennis racket, and zebra were used in the training phase, and four remaining icons of bicycle, clock, giraffe, and umbrella were introduced as new objects in the second part of the study.

For a comparison, the edge map of the photograph was extracted through kernel convolutions, which were facilitated by OpenCV, a third-party code library. Firstly, an initial Gaussian blur was computed on the entire image using a 5 × 5 kernel with SD of 0.9, to reduce the low-level noise. Canny edge detection (Canny, 1986) was then applied to the blurred image. Note that although a blurring stage is included in the Canny edge detection algorithm, another blur filter is applied beforehand as the implicit one was insufficient for the noise reduction. For the edges, the default values of 100 and 200 for the minimum and maximum of the intensity gradient respectively were used. The colours of the pixels in the edge map were inverted, so the edges are represented by black pixels and everything else by white pixels. An example can be seen in the second column of **Figure 3**. This allows the edges to be raised when printing on a swell paper that turns a picture tactile. All materials were printed on swell paper using a Zyfuse Heater.

4.3 Procedure

Participants were introduced to the study procedure, and personal data was collected including age, any visual deprivation they might have, the cause of impairment, and previous experience with braille or moon system and tactile images. Participants were asked these questions to account for any individual differences, noted by Thinus-Blanc and Gaunet (1997). Any possible relationships between such factors and performance in tactile picture comprehension, was then used for analysis. In the first part of the study, participants were familiarised with 8 out of the 12 icons. They were blindfolded and instructed to feel the eight mixed images with both hands, and were told what each icon represented, using the labels of football, car, cat, dog, elephant, person, tennis racket, and zebra. After this, the images were shuffled and participants were asked to identify each icon in 30 seconds with a 10 seconds time warning at the end. They were told whether their guess was correct, and if not, the icon was named for them after the allotted time. If participants did not name at least 7 out of 8 items correctly, the procedure was repeated until all items were identified. This was repeated up to 3 times. Finally, each participant was asked to describe their strategy for remembering the icons and given a 2-minutes break.

In the second part of the study, participants were asked to identify objects from a selection of representations, including 10 iconic and 10 edge maps. They were firstly provided with an edge map example, featuring a pigeon, to familiarise themselves with this type of representation. Following this, they were given 1 minute to identify the main objects in each representation, followed by a 10 s time warning. No feedback was given to the participants for correct or incorrect answers. Responses were marked using the labels of football, car, cat, dog, elephant, person, tennis racket, zebra, bicycle, clock, giraffe, and umbrella, or close

[4] Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., and He, K., (2018). Detectron. Retrieved from https://github.com/facebookresearch/detectron.

TABLE 1 | Percentages and Standard Deviations of correctly identified objects for each representation type for sighted and visually impaired individuals.

Sightedness	Representation type	
	Icons only	Edge maps
Sighted	57.6%, SD = 26.9	5.45%, SD = 8.01
Visually impaired	74.62%, SD = 14.14	7.76%, SD = 15.03

synonyms (e.g., time for clock, vehicle for car, etc.). If responses were vague (e.g., "animal" for giraffe), further questions were asked (e.g., "Could you be more specific on what animal you think it could be?"). Finally, participants were asked to describe their strategy for identifying the objects. The study took approximately 35 minutes.

In cases where participants were not available for a face-to-face session, the stimuli and a blindfold were posted to them, and the study was conducted through their preferred choice of video call software. Participants were instructed not to open the package prior to the study. The eight images were shuffled by themselves (or with a help of someone else) after familiarisation in the training phase, and the researcher was shown which image they were about to touch before doing so. The images for the second part of the study were premixed by the researcher before the materials were sent to the participant; the order for each participant was randomised but then corrected, so that the two representations of a given image were not presented sequentially.

4.4 Results

4.4.1 Quantitative Data

A Mixed ANOVA was performed to identify whether having a visual impairment affects the percentage of correctly identified objects; whether there is a difference in correctly identifying objects from each type of representation; and whether vision affects the percentage of correctly identified objects in each representation, as shown in **Table 1**. All relevant assumptions have been checked for, and were met including homogeneity of variances.

Results showed a significant main effect of representation type on percentage of correctly identified objects ($F(1, 30) = 222.339$, $p < 0.001$, $\eta_p^2 = .881$), suggesting that participants recognised more objects correctly if they were represented through icons rather than edge maps. A main effect of sightedness was also found ($F(1, 30) = 190.937$, $p < 0.001$, $\eta_p^2 = .864$), suggesting that visually impaired participants recognised more objects correctly, irrespective of the representation used. No interaction effect was observed ($F(1, 30) = 3.394$, $p = 0.075$), meaning that sightedness did not affect which representation participants recognised more objects from.

A paired sample t-test was used to identify any differences in the percentage of correctly identified objects that were learned during the training phase or not. A percentage of 71.57% (*SD* = 25.93) learned icons were correctly recognised, while 33.87% (*SD* = 29.51) of new icons were correctly recognised. A significant difference between the learned and new icons was found ($t(30) = 8.189$, $p < 0.001$), suggesting that the icons learned beforehand are more likely to be recognised than new icons that are new. Pearson's 2-tailed correlation revealed that participants with more correctly identified learned objects also correctly identified more new objects, ($\rho = 0.579$, $p = 0.001$). It must also be noted that no significant correlation was found between participants' age and percentage of correctly identified objects ($\rho = 0.357$, $p = 0.045$). Furthermore, no significant correlation was found between the number of correctly recognised learned icons and correctly identified edge map objects ($\rho = 0.224$, $p = 0.083$).

4.4.2 Qualitative Data

Recognising objects in tactile images at a high rate, although useful, is by no means the end of comprehension. The explanations and narratives that people construct about the content is important, although due to the nature of qualitative data, it is harder to interpret. Even though story building was not directly instructed, participants did construct explanatory narratives.

After recognising icons of a person and a tennis racket, participants assumed that the picture was of a person playing tennis (see **Figure 4A**). In the case of two people playing football, the story was "playing football". Some pictures were more complex and harder to interpret. One such photograph showed a cyclist next to a car on a road. Our algorithm recognised all the parts, but displayed them as non-overlapping icons for a person, a bicycle, and a car. The requirement for icons not to overlap confused the narratives: one participant said "a person is trying to decide whether to commute by a car or by bike". Another felt it was "a person crossing the road with his bike" or "a person with his car and bike parked". **Figure 4B** provides other interpretations. This shows the innate ambiguity in some of the pictures.

However, being able to make a plausible narrative at all, is arguably preferable to making no narrative at all, and participants rarely gave such commentaries when presented with edge maps. This may be expected if recognition is a perquisite to any explanation. (If nothing is recognised, there is nothing to explain.) Rather, as in **Figure 4**, participants commented on characteristics of the tactile image that are extraneous to its semantic content, such as the textural experience it gave.

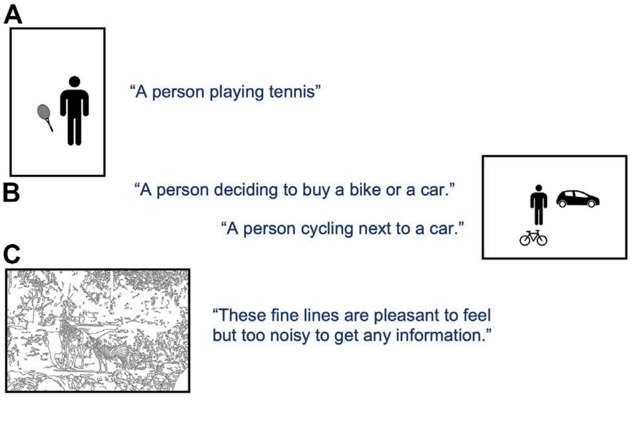
A
"A person playing tennis"
B
"A person deciding to buy a bike or a car."
"A person cycling next to a car."
C
"These fine lines are pleasant to feel but too noisy to get any information."

FIGURE 4 | Comments by different participants for some of these representations.

5 DISCUSSION

The combination of Mask R-CNN (He et al., 2017) and PiCANet (Liu et al., 2020) made it possible to automatically detect, identify, and localise salient objects. The icon-based tactile representation worked particularly well among Visually Impaired participants: 74.62% for icons and 7.76% for edge maps. Many participants enjoyed grasping some information about photographs through simple iconic representations; in contrast, feeling edge maps led to cognitive overload. Moreover, participants were able to construct relevant narratives from the icon images, but rarely from edge maps. Not only were objects more easily identifiable through icons, but it also brought the story within photographs to life.

Icons and edge maps are not the only choices available when processing photos into tactile images. When artists draw, it is common for them to deliberately highlight the most salient aspects of a scene, providing an idea of a possible alternative route for automation. The literature provides many methods for making strokes (Hertzmann et al., 2002; Lang and Alexa, 2015), including the contemporary use of neural networks (Li et al., 2019b,a). However, the question of where to place strokes is crucial; some reasonable solutions exist for 3D models (DeCarlo et al., 2003) but the case for photographs remains open. Therefore, the use of icons is a direct and available route to demonstrate semantic content. Our particular choice of style for icons was based on previous psychology research works, i.e., 2D representations are easier to recognise than 3D, and is the preferred method for recognising objects in three-quarters front view (Bartram, 1974; Palmer, 1981; Klatzky et al., 1985; Biederman, 1987; Lederman et al., 1990; Eriksson, 1999).

The importance of icon size should not be overlooked when developing images. While size variation of icon was notified in the information sheet, most participants did not remember this and were often surprised by the change. This resulted in low confidence and some incorrect responses. For example, many participants identified the elephant icons as dogs because icons of elephants were small in the training phase. If participants used the icon images for longer, then confidence and judgment may return.

One particularly interesting issue that arose from scale, was confusion between perspective scaling and actual object size. Some participants described the football iconic picture (**Figure 3D**) as football being played by either “a man and a child” or “two people at different distances.” Congenitally blind participants tended to describe icons as big or small, rather than the corresponding objects as being close or far.

Icon size also affected the textural features of an icon. The tennis racket was sometimes identified as a lollipop or a balloon; and the zebra icon as a horse or a dog. Tactile resolution for the skin is approximately 2.5 mm (Sherrick and Craig, 1982). In other words, when two points are closer, they tend to feel like one point. It could be assumed that these icons were not recognised correctly because of low touch sensitivity as noted by Manning and Tremblay (2006); Tremblay and Master (2015), but incorrect responses were made by participants of varying tactual experiences which agrees with research works by Legge et al. (2008) and Heller (1989). Thus, the implementation of size criteria of icons might be necessary to improve intelligibility.

For the strategy of remembering the icons, participants often looked for specific features first, like texture or distinctive parts, rather than outlines. Several participants commented that visualising an icon as a whole was difficult. For example, a cat was remembered by its curved tail, a zebra by their stripes, an elephant by its trunk, a car by their wheels, and a tennis racket by their strings. The umbrella icon was sometimes identified as a cat, because the curved tail felt like the curved handle. The umbrella icon was not included in the training phase of the experiment, so it had to be learned. Again, greater familiarity may mitigate such issues, but this is an open question.

Some participants developed a strategy for identifying icons of animals, by remembering which direction these fixed icons were facing in the training phase of the study. That is, all animal icons were facing the right, except the zebra. This strategy was particularly useful when size of icons changed drastically and some features were no longer touchable, e.g., after the icon of a zebra turned very small, its their stripes were no longer palpable to some participants, but it was still correctly identified due to direction. One participant wondered whether these icons would still be recognisable if they were mirrored. Since icons were primarily remembered by their distinctive features rather than as a whole, it would be interesting to investigate whether icons with varying poses would be recognisable should the distinctive features remain fixed (i.e., tail of a cat stays unchanged while the body could vary to match the pose on the photograph).

With the effect of reduced performance in object identification, when icons changed size after training, we could assume that it would be harder to recognise icons that change their poses accordingly. However, the first part of the experiment consisted of fixed sized icons only, thus it is unclear whether the outcome would have been different if participants were trained to recognise icons in different sizes.

It is prominently noticeable that our icon representations displayed only some of the foreground of the picture and none of the background. Also, our algorithm does not allow icon overlap. Both of these characteristics contributed to the diversity of narratives about the tactile images. Biederman (1981) has presented some results in object identification and found that produced accuracy varies when objects are presented on different types of background. With many more research studies on gist of the scene, such as Davenport and Potter (2004) or Munneke et al. (2013), it is clear that objects are identified more accurately when the background scene is consistent (e.g., a sandcastle on a beach instead of a sandcastle on a road). Future work would be needed to determine whether this effect is similar in touch senses, alongside an investigation of whether objects correctly identified by facilitation of background information, leads to a better contextual understanding. How to present background information tactually and intelligibly would be an important avenue for future research.

Another potential area for future work is having the neural network learn the thresholds discussed in the rendering phase. This can be achieved with transfer learning, so the model also includes the numeric threshold value as an output of the final layer. The difficulty with this is finding, or creating data that has adequate thresholds for many scenes.

With some future work listed and no particular focus on the best neural networks to use, our algorithm has a potential to bring a new accessibility feature to the visually impaired community - a feature that automatically communicates information about images, saving time and reducing costs of tactile image production.

6 CONCLUSION

Our approach for rendering tactile images by detecting important objects, communicates content in such a way that people can construct a meaningful narrative of the visual image. Overall, participants reported that edge maps have too much noise to be useful.

We can display foreground icons only in a fixed pose and in a non-overlapping fashion. Considerable work is needed to discover how to produce better icons automatically. One prominent gap in our iconic representations is that it has no background information. Such inclusion poses a considerable challenge, which brings a research opportunity for future work.

Nonetheless, we have provided the first known approach that is capable of communicating some content *via* touch; a small step technically perhaps, but with millions world-wide having little or no access to photos, this is a small step with a potentially large impact.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion 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 Simon Jones, Department of Computer Science, University of Bath, United Kingdom. Participation in the study was entirely voluntary and electronic written informed consent was provided to participate in this study.

AUTHOR CONTRIBUTIONS

PH created the original idea for this project. PN first implemented the algorithm and KP made further edits. MP and KP designed the experiments and KP organised the database of participants. KP and EK together conducted the qualitative and quantitative analyses. KP wrote the first draft of the manuscript. EK, PN, PH, MP, and KP wrote sections of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

FUNDING

Research by KP is supported in part by the UKRI EPSRC Centre for Doctoral Training in Digital Entertainment (CDE), EP/L016540/1. Research by MP is partly funded by CAMERA 2.0, the UKRI Centre for the Analysis of Motion, Entertainment Research, and Applications (EP/T014865/1).

ACKNOWLEDGMENTS

We would like to say many thanks to participants in the UK and Greece for volunteering in our study. Their encouraging co-operation and comments made this paper possible. We would also like to say many thanks to Sarah Pool for language editing. Her timely assistance was greatly appreciated. It is also acknowledged that material from Petar's dissertation was used in this paper (Nedelev, 2019). The rest of the images used in a study are available on request.

REFERENCES

Bartram, D. (1974). The Role of Visual and Semantic Codes in Object Naming. *Cogn. Psychol.* 6, 325–356. doi:10.1016/0010-0285(74)90016-4

Biederman, I. (1981). Do background Depth Gradients Facilitate Object Identification? *Perception* 10, 573–578.

Biederman, I. (1987). Recognition-by-components: a Theory of Human Image Understanding. *Psychol. Rev.* 94, 115.

Borji, A., Cheng, M.-M., Jiang, H., and Li, J. (2015). Salient Object Detection: A Benchmark. *IEEE Trans. Image Process.* 24, 5706–5722. doi:10.1109/tip.2015.2487833

Brown, D. J., and Proulx, M. J. (2016). Audio-vision Substitution for Blind Individuals: Addressing Human Information Processing Capacity Limitations. *IEEE J. Selected Top. Signal Process.* 10, 924–931. doi:10.1109/JSTSP.2016.2543678

Büchel, C., Price, C., Frackowiak, R., and Friston, K. (1998). Different Activation Patterns in the Visual Cortex of Late and Congenitally Blind Subjects. *Brain a J. Neurol.* 121, 409–419.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Machine Intelligence PAMI-* 8, 679–698. doi:10.1109/TPAMI.1986.4767851

Cantoni, V., Lombardi, L., Setti, A., Gyoshev, S., Karastoyanov, D., and Stoimenov, N. (2018). "Art Masterpieces Accessibility for Blind and Visually Impaired People," in International Conference on Computers Helping People with Special Needs, Linz, Austria, July 11-13, 2018 (Springer), 267–274.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end Object Detection with Transformers," in *European Conference on Computer Vision* (Paris: Springer), 213–229.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Davenport, J. L., and Potter, M. C. (2004). Scene Consistency in Object and Background Perception. *Psychol. Sci.* 15, 559–564. doi:10.1111/j.0956-7976.2004.00719.x

DeCarlo, D., Finkelstein, A., Rusinkiewicz, S., and Santella, A. (2003). Suggestive Contours for Conveying Shape. *ACM Trans. Graphics (Proceedings SIGGRAPH)* 22, 848–855.

Eriksson, Y. (1999). "How to Make Tactile Pictures Understandable to the Blind Reader," in Proceedings of the 65th IFLA Council and General Conference, Bangkok, Thailand, August 20-28, 1999.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behav. Res. Methods* 41, 1149–1160. doi:10.3758/BRM.41.4.1149

Gallace, A., and Spence, C. (2009). The Cognitive and Neural Correlates of Tactile Memory. *Psychol. Bull.* 135, 380. doi:10.1037/a0015325

Hayhoe, S. (2013). Expanding Our Vision of Museum Education and Perception: An Analysis of Three Case Studies of Independent Blind Arts Learners. *Harv. Educ. Rev.* 83, 67–86.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 29, 2017 (Institute of Electrical and Electronics Engineers IEEE), 2980–2988. doi:10.1109/ICCV.2017.322

Heller, M. A., and Kennedy, J. M. (1990). Perspective Taking, Pictures, and the Blind. *Perception & Psychophysics* 48, 459–466. doi:10.3758/BF03211590

Heller, M. A. (1989). Picture and Pattern Perception in the Sighted and the Blind: The Advantage of the Late Blind. *Perception* 18, 379–389. doi:10.1068/p180379

Hertzmann, A., Oliver, N., Curless, B., and Seitz, S. M. (2002). Curve Analogies. *Rendering Tech.* 2002, 13th. doi:10.5555/581896.581926

Kennedy, J. M. (1993). *Drawing & the Blind: Pictures to Touch*. Yale University Press.

Klatzky, R. L., Lederman, S. J., and Metzger, V. A. (1985). Identifying Objects by Touch: An "Expert System". *Perception & Psychophysics* 37, 299–302. doi:10.3758/bf03211351

Klatzky, R. L., Loomis, J. M., Lederman, S. J., Wake, H., and Fujita, N. (1993). Haptic Identification of Objects and Their Depictions. *Perception & Psychophysics* 54, 170–178. doi:10.3758/BF03211752

Kurze, M. (1997). "Rendering Drawings for Interactive Haptic Perception," in Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Atlanta Georgia USA, March 22-27, 1997 (Berlin: Association for Computing Machinery), 423–430. doi:10.1145/258549.258826

Lang, K., and Alexa, M. (2015). "The Markov Pen: Online Synthesis of Free-Hand Drawing Styles," in Proceedings of the Workshop on Non-Photorealistic Animation and Rendering (Eurographics Association), Istanbul Turkey, June 20-22, 2015 (Berlin: NPAR), 203–215.

Lederman, S. J., Klatzky, R. L., Chataway, C., and Summers, C. D. (1990). Visual Mediation and the Haptic Recognition of Two-Dimensional Pictures of Common Objects. *Perception & Psychophysics* 47, 54–64. doi:10.3758/bf03208164

Legge, G. E., Madison, C., Vaughn, B. N., Cheong, A. M., and Miller, J. C. (2008). Retention of High Tactile Acuity throughout the Life Span in Blindness. *Percept Psychophys* 70, 1471–1488.

Li, M., Lin, Z., ech, R. M., Yumer, E., and Ramanan, D. (2019a). *Photo-Sketching: Inferring Contour Drawings from Images*. Pittsburgh: WACV.

Li, Y., Fang, C., Hertzmann, A., Shechtman, E., and Yang, M.-H. (2019b). "Im2Pencil: Controllable Pencil Illustration from Photographs," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 15-20, 2019, 1525–1534.

Lin, A., Scheller, M., Feng, F., Proulx, M. J., and Metatla, O. (2021). *Feeling Colours: Crossmodal Correspondences between Tangible 3D Objects, Colours And Emotions (Association for Computing Machinery)*. Bristol.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft Coco: Common Objects in Context," in European conference on computer vision, Zurich, Switzerland, September 6-12, 2014 (Springer), 740–755.

Liu, N., Han, J., and Yang, M.-H. (2020). PiCANet: Pixel-Wise Contextual Attention Learning for Accurate Saliency Detection. *IEEE Trans. Image Process.* 29, 6438–6451. doi:10.1109/TIP.2020.2988568

Manning, H., and Tremblay, F. (2006). Age Differences in Tactile Pattern Recognition at the Fingertip. *Somatosensory Mot. Res.* 23, 147–155. doi:10.1080/08990220601093460

Melloni, L., van Leeuwen, S., Alink, A., and Müller, N. G. (2012). Interaction between Bottom-Up Saliency and Top-Down Control: How Saliency Maps Are Created in the Human Brain. *Cereb. Cortex* 22, 2943–2952. doi:10.1093/cercor/bhr384

Munneke, J., Brentari, V., and Peelen, M. (2013). The Influence of Scene Context on Object Recognition Is Independent of Attentional Focus. *Front. Psychol.* 4, 552. doi:10.3389/fpsyg.2013.00552

Nedelev, P. (2019). *Photos for the Visually Impaired*. Bath: Bachelor's thesis, University of Bath.

Oliva, A., and Torralba, A. (2006). "Chapter 2 Building the Gist of a Scene: the Role of Global Image Features in Recognition," in *Visual Perceptionof Progress in Brain Research* (Elsevier), Vol. 155, 23–36. doi:10.1016/S0079-6123(06)55002-2

Palmer, S. (1981). Canonical Perspective and the Perception of Objects. *Attention Perform.* 135–151.

Panotopoulou, A., Zhang, X., Qiu, T., Yang, X.-D., and Whiting, E. (2020). Tactile Line Drawings for Improved Shape Understanding in Blind and Visually Impaired Users. *ACM Trans. Graph.* 39. doi:10.1145/3386569.3392388

Pascolini, D., and Mariotti, S. P. (2011). Global Estimates of Visual Impairment: 2010. *Br. J. Ophthalmol.* 96, 614–618. doi:10.1136/bjophthalmol-2011-300539

Reichinger, A., Carrizosa, H. G., Wood, J., Schröder, S., Löw, C., Luidolt, L. R., et al. (2018). Pictures in Your Mind: Using Interactive Gesture-Controlled Reliefs to Explore Art. *ACM Trans. Accessible Comput. (Taccess)* 11, 1–39. doi:10.1145/3155286

Romera-Paredes, B., and Torr, P. H. S. (2016). "Recurrent Instance Segmentation," in European conference on computer vision, Amsterdam, Netherlands, October 11-14, 2016 (Springer), 312–329.

Rossi, D., Maglione, A. G., Modica, E., Flumeri, G. D., Venuti, I., Brizi, A., et al. (2017). "An Eye Tracking index for the Salience Estimation in Visual Stimuli," in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Jeju, Korea (South), July 11-15, 2017 (Rome: EMBC), 4483–4486. doi:10.1109/EMBC.2017.8037852

Salvador, A., Bellver, M., Campos, V., Baradad, M., Marques, F., Torres, J., et al. (2017). *Recurrent Neural Networks for Semantic Instance Segmentation*. Barcelona: arXiv preprint arXiv:1712.00617.

Sherrick, C. E., and Craig, J. C. (1982). The Psychophysics of Touch. *Tactual perception: A sourcebook* 55–81.

Stangl, A., Morris, M. R., and Gurari, D. (2020). ""Person, Shoes, Tree. Is the Person Naked?" what People with Vision Impairments Want in Image Descriptions," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu HI USA, April 25-30, 2020.

Thinus-Blanc, C., and Gaunet, F. (1997). Representation of Space in Blind Persons: Vision as a Spatial Sense? *Psychol. Bull.* 121, 20.

Tremblay, F., and Master, S. (2015). Touch in Aging. *Scholarpedia* 10, 9935. doi:10.4249/scholarpedia.9935

Way, T. P., and Barner, K. E. (1997). Automatic Visual to Tactile Translation. I. Human Factors, Access Methods and Image Manipulation. *IEEE Trans. Rehabil. Eng.* 5, 81–94.

Wu, S., Wieland, J., Farivar, O., and Schiller, J. (2017). "Automatic Alt-Text: Computer-Generated Image Descriptions for Blind Users on a Social Network Service," in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Portland Oregon USA, February-March 25-1, 2017, 1180–1192.

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.