Face Recognition Assistant for People with Visual Impairments

MOHAMMAD KIANPISHEH, University of Toronto, Toronto, ON, Canada
FRANKLIN MINGZHE LI, University of Toronto, Toronto, ON, Canada
KHAI N. TRUONG, University of Toronto, Toronto, ON, Canada

[1] Although there are many face recognition systems to help individuals with visual impairments (VIPs) recognize other people, almost all require a database with the pictures and names of the people who should be tracked. These solutions would not be able to help VIPs recognize people they might not know well. In this work, we investigate the requirements and challenges that must be addressed in the design of a face recognition system for helping VIPs recognize people with whom they have weak-ties. We first conducted a formative study with eight visually impaired people. Using insights learned from the formative study, we developed a research prototype that runs on a mobile phone worn around the user's neck. The developed prototype is a wearable face recognition system that opportunistically captures and stores undistorted face images and contextual information about the user's interaction with each person to a database, without the user intervention, as she interacts with new people. We then used this prototype application as a technology probe—asking VIP participants to use the device in a realistic scenario in which they meet and re-encounter several new people. We analyze and report feedback collected from VIPs about the design and use of such a service.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing;**

KEYWORDS

Visually impaired people, face recognition, weak-ties

ACM Reference format:

1 INTRODUCTION

Not being able to recognize other people is one of the main challenges that people with visual impairments (abbreviated as VIPs) face in their daily life. For example, a VIP might enter a busy elevator and encounter a neighbor with whom she had a small chit-chat at a party last week for the first time. In this situation, she might not know this neighbor is in the elevator unless the other person sees her and greets her first. If the neighbor were to greet her, the VIP might struggle to recognize the neighbor, because she does not know the neighbor well. Situations like this can be awkward because either the VIP must pretend to recognize the other person or the neighbor must explain who he is and how they know each other. Difficulties with knowing when they have encountered someone they have met previously and recognizing those individuals can impede VIPs' ability to interact with people they only know casually or in a limited capacity—

This work is supported by NSERC.

Authors' addresses: Mohammad Kianpisheh, University of Toronto, Toronto, ON, Canada, kian@dgp.toronto.edu; Franklin Mingzhe Li, University of Toronto, Toronto, ON, Canada, franklin.li@mail.utoronto.ca; Khai N. Truong, University of Toronto, Toronto, ON, Canada, khai@cs.toronto.

https://doi.org/10.1145/1234

individuals often referred to as weak-ties in the psychosocial literature [10]. While strong-ties (the people whom a person knows well, such as family members and close friends) have a significant impact on one's well-being and emotional health [44,45], being ensconced tightly with strong-ties only may not be optimal. Often a person and her strong-ties share similar values, knowledge, and connections; however, weak-ties are valuable in that they serve as a door to the novel sources of information [10]. Studies also have demonstrated the positive impact of weak-ties on people's emotional health: for example, people with more weak-ties are healthier and happier and experience less anxiety and stress than those with fewer weak-ties [11].

Researchers have explored designing face recognition assistance systems help VIPs recognize people [14, 20, 33, 24]. However, none of the previous work has taken into account the need for weak-ties recognition. For example, all of the current face recognition systems were designed to associate each face with the name of a person from a defined set of people that the user knows well. However, weak-ties are typically individuals who fall outside of this set of people, such as acquaintances whom the user see infrequently or someone she only knows casually (e.g., a cashier at the grocery store). They may only have pictures of people in their social network for most of the face recognition systems to create a database of faces to recognize. They are not likely to have pictures of people whom they know only in a limited capacity to enroll into a face recognition assistance system. Moreover, the VIPs might not even know the other person's name. So how would they enroll such a person into an aide? What information should the system provide about the other person, if a name is not available, to help the user recognize who she has encountered?

In this research, we studied how to design a face recognition system that can help VIPs in recognizing acquaintances and people they do not know well, and explored the requirements and design challenges that such a system must address in real-life like scenarios. To this end, first, we conducted a formative interview study to better understand the VIPs' challenges with recognizing their weak-tie connections and the features that should be included in such a system. Then, we used formative insights to develop a face recognition research prototype that runs on a mobile phone worn around the user's neck to assist VIPs in recognizing people with whom they are not necessarily closely connected. When a VIP interacts with others, the developed prototype determines if the face is new or exists in a database of individuals whom the VIP has interacted with previously. If the face is new, the system opportunistically captures a few undistorted face images of the person's face and adds them to the database. To help VIPs recognize their weak-ties, the system assigns each person in the database with information that might help the user to recognize this individual in future encounters. Furthermore, VIPs can complement the system information by providing their description of the person in the form of an audio message.

We used this prototype application as a technology probe—asking 10 VIP participants to use the device in a realistic real-life scenario in which they would meet and interact with different people for the first time and then again—to elicit feedback from VIPs about the design and use of such a service when meeting and re-encountering new people. We tried to gain a greater understanding about the usefulness of the information clues that our participants considered important in the formative study, and to learn about the system requirements and challenges involved with collecting the relevant information without the user intervention. Findings from this study will provide insights about how to study and develop assistive technologies that will help VIPs recognize their weak-ties in real-life scenarios and uncover the challenges associated with real-time interaction with such an assistant system.

2 RELATED WORK

There are two streams of research related to this work. One stream is face recognition systems designed to assist VIPs in recognizing faces belonging to people who have been enrolled in the system; the other focuses on systems which provide information to remind the user about the person whom she is interacting with.

2.1 Memory augmentation for name recall

Memory augmentation has been considered as the killer application for wearable computers to the general masses [20], and it has been an area of high interest in recent years. Here, we focus on memory augmentation systems designed to help users in recognizing (recalling) people they have encountered before. For example, Fenwick *et al.* [33] developed a mobile application that associates people with the time and location that are more likely to be seen. Their system database is composed of people that are already in the user's contact. The user can also manually create new entities and provide additional metadata for each person. The researchers conducted pre-pilot studies at each stage of the system development, but no user study has been completed with the developed application.

Researchers have also designed memory augmentation systems on wearable devices. Utsumi et al. [23] developed a wearable face recognition system performing in real-time where face detection, tracking, and recognition are performed on a computer and recognition results are superimposed on the video displayed on a head-mounted display (HMD). Their evaluation with public and custom datasets shows that the developed system can perform face recognition at the speed of 668 ms and an accuracy of 93.3%. Wang *et al.* [21] designed a face recognition system using an AR eyewear to help people who are diagnosed with prosopagnosia or inability to recognize faces. In their system, the mounted camera on the eyewear sends video frames to a smartphone where face recognition is performed. After the face recognition is completed, each detected face in the scene will be augmented by the name and her relation to the user. No user study has been reported; however, their evaluations with collected and public dataset show that the trained face recognizer can achieve an average accuracy of 99% when a frontal face is in the viewpoint of the camera. Researchers have also used contextual information to help users to recognize people they have met before. For example, Iwamura *et al.* [22] designed a memory assistant system that shows to the user previous encounters with a person that she is currently looking at. In addition to videos, the system displays the times and the location of their encounters. Through an online questionnaire, they learned that respondents considered places and events of previous encounters as useful clues for reminding them of people they have met before. Researchers have also proposed systems that associate people's face with information about them that is available on social media. For example, Smart Glasses [24] is a prototype wearable AR display that augments the people's face in front of the user with their social media information. Kurze and Roselius implemented the Smart Glasses as an Android application that runs on smartphones; however, no evaluation is done for the proposed system.

The memory augmentation systems mentioned above are typically evaluated on datasets pre-collected in controlled environments. No user study has been done to evaluate the performance of these developed wearable face recognition systems in real-life scenarios and gather the user's feedback about the provided memory cues by the system. Also, none of these prior works target visually impaired people. Unlike sighted people, VIPs can not benefit from visual cues, and they may need different requirements in the developed assistive technology.

2.2 Face recognition assistive systems for visually impaired people

For almost two decades, researchers have designed assistive technologies for VIPs to assist them in recognizing people. Kramer et al. [15] developed a face recognition application for smartphones to aid VIPs in recognizing people. In their system, pictures taken by the smartphone are sent to a server for face recognition. To create the system database, ten people were enrolled with pictures of their faces from 15 different positions (i.e., different angles, and distances). They tested their system in a conference setting where subjects and the experimenter sat around a table, and the experimenter takes pictures from all subjects using a smartphone when they are sitting in different chairs around the table. They showed that

the proposed method could detect and recognize faces when the subject is looking away from the camera at a 40-degree angle. Otherwise, the system could not detect the subject's face.

Similar to our work, Chaudhry and Chandra [18] previously explored the use of a smartphone for face recognition. Their system has two modes: offline and online. In the offline mode, the recognition computation takes place on the smartphone itself, while the frames are sent to the server for face recognition in the online mode. Their system also has an enrollment mode in which the system stops to identify faces, and the next detected face will be considered as a new identity, and the user can add that into the database; however, no more details are provided for the enrollment mode. The work was evaluated using experiments on a custom video dataset recorded by a smartphone from four individuals which the proposed method achieved an accuracy of 52% to 70% for different test videos. As they reported, their highest accuracy (70%) achieved when the test video was taken in a well-lit condition with the subject looking directly to the camera; this shows the adverse effect of poor illumination and face angle on face recognition performance.

Other devices, such as eyeglasses and smartwatches, have also been used to help VIPs recognize the faces of people they know. For example, Krishna *et al.* [17] designed sunglasses with a camera on its nose bridge to capture images of people that the VIP interacts with. Their system then uses Principal Components Analysis (PCA) for face recognition. To make their face recognition algorithm robust to illumination and head-pose change, they created a custom dataset composed of 450 images of 10 individuals with various illumination level and head pose. They used different methods for face recognition and showed that the Linear Discriminant Analysis (LDA) method achieves the best accuracy of more than 90%. OrCam [46] is also a commercial camera mounted on the user's eyeglasses which announces the subject's name when she appears in front of the VIP. To form the system database, OrCam requires the VIP to take pictures of people whom she wants to recognize in future encounters; however, there is no user study reported in the literature about OrCam.

Laurindo et al. [42] developed a face recognition system running on a smartwatch. They attached a camera to the smartwatch wristband to enable the user to scan her surroundings. When the system detects a face, the user must then hold the camera still for several seconds for the system to recognize who from a database of images that person is. The system database is populated with five pictures of everyone that the user wants to recognize. In their experiments, five blindfolded participants tried to locate three people who are standing in a room at a short distance. Laurindo et al. [42] reported a success rate of 0.83 while it took the participants on average between 1 min to 3 min to locate people in the room for different rounds of the study. They used the K-nearest neighbor (KNN) algorithm to classify faces in which a face will be assigned to a new class if its distance to all samples in the database is larger than a threshold. The user can also associate audio to the new detected class. However, no user study and details have been reported for this part. More recently, depth cameras have been explored in place of normal RGB cameras for face recognition. For example, Laurindo et al. [16] used a Microsoft Kinect sensor (worn by the user on a helmet) to perform the face detection. Detected faces were then classified using the KNN algorithm. Once the system detects a familiar face, a 3-D sound is presented to the user through a bone conduction headphone to virtualize the location of the recognized person. They tested the system on a custom dataset collected from 15 subjects (20 videos for each subject) captured using a Kinect. In their experiments, five visually impaired people tried to locate and recognize three subjects standing in the room. After 15 rounds of experiments, they reported a success rate of 53%. They argue that the bright background in the room can affect the system performance; this illustrates some of the challenges of using wearable devices for face recognition.

All of the works mentioned above have been evaluated under controlled environment in which participants try to recognize subjects who do not move and are standing in the testing environment. Furthermore, the proposed methods database is composed of pre-collected identities, which a sighted individual would collect pictures or record videos of subjects whom the participant wants to recognize. Also, in most of the previous works, the system database is populated by face images of a fixed number of

peoples. Although Chaudhry and Chandra [18], and Laurindo *et al.* [42] considered the capability of adding new subjects to the database, no evaluation or user study is provided to test this application behaviour.

One exception is the system proposed by Zhao *et al.* [14]. Using Facebook's face recognition algorithm [43], they developed a bot which provides the VIPs with information such as people's identity, face expressions and attributes through the user's smartphone. In their system, the VIP takes a picture of the person she wants to recognize. Their system database is composed of the Facebook friends of the user who have the Facebook tag suggestions option turned on. Then the bot sends the picture to a remote server for recognition. The server then sends back the results to the smartphone. They conducted a seven days diary study in which six VIPs tested their system in their daily life. Their follow-up interview study showed that low-vision participants found their system to be helpful in assisting them to recognize other people. They reported difficulty in aiming the camera as one of the major challenges that their participants faced in using their system.

None of the previous works were concerned with weak-ties nor shed light on how to design a system to recognize people who are not close-ties with the user. Because they do not examine that context, they do not report on information such as what types of information would be useful to VIPs, and the system requirements that need to be included in such a system.

3 FORMATIVE STUDY

We conducted a formative interview study to better understand the importance of recognizing weak-ties. Moreover, we asked the participants about the cues that might help them to recognize their weak-ties.

Table 1. Demographics of the formative study participants

Participant	Age/Sex	Occupation	Vision condition
P1	35/M	University professor	Blind since he was 4
P2	49/M	HR & EE coordinator	Blind since he was 2
P3	32/M	Tele marketer	Low vision; cannot recognize faces
P4	53/M	unemployed	Blind since he was 9
P5	56/F	unemployed	Low vison; cannot recognize faces
P6	60/F	unemployed	Ultra low vision; only can see shapes
P7	21/F	Undergraduate student	Low vision; cannot recognize faces
P8	51/M	unemployed	Blind since he was 9

3.1 Method

We interviewed eight legally blind participants (3 female) between 21 to 59 years of age (mean=44.6). Table 1 shows our participants' demographic information, including their vision condition. We used a semi-structured interview format, asking VIPs about the challenges they face in recognizing people they do not

know well and cues that might help them to do so. The interviews took on average around 51 minutes, ranging from 33 to 73 minutes. We compensated all participants $30 CAD for their involvement in the study. After transcribing all of the interviews, two of the researchers (the first and second authors) coded one of the interviews separately using open coding. They then jointly reviewed and discussed the coded interview and the categories until they came to an agreement. After that, one of the researchers (the first author) coded the remaining interviews based on the agreed categories.

3.2 Findings

3.2.1 The importance of recognizing weak-ties. All participants acknowledge the importance of social interactions with weak-tie connections. An important problem for VIPs is to not appear rude to others. For example, P6 mentioned: *"It is important to acknowledge people even if I don't know their name, it doesn't matter if I can't see them .... they might have helped me before but I can't thank them."* Unfortunately, not being able to see the other person limits the VIP's ability to extend previous interactions into the most recent encounter: *"Every time I meet them, it's just hello again, like a starting point every time I meet them."* Three participants also mentioned they would feel secure if they could know when there are familiar people around them or not. For example, *"sometimes I am with my family downtown shopping, and somebody come and give a hug; it sounds like they know me but I don't know them; ... Some of them maybe are mental people (sic) or drug addict I can't tell. You know it's dangerous in the downtown area"* (P2). Two participants also mentioned that interacting with weak-ties exposes them to new opportunities for example when they are seeking a job. "When I was at the gym, I have been actively trying to network and find a job. Going to the networking environment by myself was super challenging. If I see somebody I might have given my card to, and they have given their name, instead of reconnecting with them, I don't recognize them because I only have met them once in that networking environment; I wouldn't see them every day" (P6).

3.2.2 Helpful clues for recognition. ***Voice as the primary clue for recognition:*** All participants considered voice as the most crucial clue in recognizing people; however, they mentioned that they might have trouble recognizing their weak-ties' voice. As P3 described *"The voice is important. If I can recognize the voice, I would say okay, that is someone I know .... unless this is someone I don't deal with often. There was someone I went to a vocation program with a couple of weeks ago, and he saw me in the mall "oh do you remember me.". It took me maybe a minute or two to recollect who he is."* Our participants also mentioned that they have difficulty in recognizing people's voice in crowded areas. *"In the crowd, it is hard to recognize people from their voices; like in funerals and weddings people come to talk to me, and I know they are family, but I can't remember who they are." (P5)*

Contextual cues: All participants thought that location was an effective cue that triggers their memory when trying to recall who is the person that they are talking to. Three participants mentioned that it is hard for them to recognize people when they are in places where VIPs do not expect them to be. As P5 said *"Some people's voice is very distinct, but if I see them in a different building then I wouldn't be able to recognize them. If I expect a person in a certain building, certain meeting then I know easier, but it is hard if they are in a totally different unexpected place."*

Name is not always helpful: Although people's names can be an effective way to help VIPs recognize people, our participants mentioned scenarios in which they cannot recognize people when given their name. One reason for this is that VIPs do not feel comfortable asking for people's name (P2, P3, P4, and P5). *"Sometimes they [my co-workers] help me but too bad I can't trace them back. Sometimes I ask their name, but I can't ask people's name on this floor all the time"* (P2). Besides feeling uncomfortable, sometimes VIPs do not even want to know their acquaintance's names. As P5 mentioned, *"I don't have any reason to go and ask the security guard's name."* Furthermore, two participants (P5, and P6) mentioned that they might not recognize people even when they are given the name. *"In my culture, there are zillions of Mala; so they have to say Mala from where. You need a reference, not just the name"* (P5).

Physical attributes: Physical attributes and people's clothing are clues that participants with low-vision (P3, P5, and P7) considered being crucial in recognizing people. *"I use general features like hairstyle and what they wear"* (P5). However, sometimes people with low-vision struggle to identify people in this manner because of changes in clothing as P5 mentioned: *"When I go to a conference and people change their clothes at night for a bar or something I can't recognize them. They say but you saw me today and I'd say yes but you changed your clothes."* It is also challenging to distinguish people whose clothes are not distinctive. For example, *"At the grocery store, I know the girls [who work there], and I see them every week…. I won't recognize them because all of them wear the same uniform"* (P3). P7 also considers lighting to be an important factor in being able to recognize people by their appearance and she finds it challenging to do so when lighting conditions change. *"It has to be a bright room or something like that [usually]."*

Other recognition cues: Our participants mentioned other additional information that can help them in recognizing people. Some participants find references to their previous interaction with people helpful like *"danced with me at the wedding"* (P5) or the content and purpose of their previous conversation. People's career "*Cashier at Metro "* and profession *"the guy in the band"* (P5) are other clues that all participants find useful in recognizing people. P2 and P7 also considered their connecting point with other people helpful. *"I need to know our connecting point (…) who is our mutual friend."* (P2). P8 also considered scent as an important clue as he described: *"There is a Wheel-Trans [a transportation service for people with disabilities] driver who usually chews betel nuts and leaves, so I can recognize him."* In general, for people they do know not well, participants often want to know how they know the person.

3.2.3 Helpful clues for recognition of close-ties. Although people's voice as a clue may not be effective enough to recognize people that VIP do not interact with regularly, our participants mentioned that they recognize their close-ties like family members and close friends merely by their voice and the way they talk. As P6 described: *"My family members, they just have to say hello and I know who it is, even if they don't have very distinctive voice…. It is because of the amount of time I talked to them and I am used to their pronunciation and the tone of their voice [and] how they speak. So it is voice and language and everything that goes along with that."* In general, as VIPs spend more time with their close-ties, they can acquire a lot of information which helps them to identify their family members and close friends. For example, P7 can identify people whom she knows well by the way they walk. *"If I know you, and I've seen you for enough time, I know how you walk (….) I was a ballet dancer for fifteen years and I learned to walk in a certain way [and] hold my posture in a certain way and there is a certain movement that goes along with. So I always pay attention to how other people walk; someone walks stiffly and someone bops his head as he walks; then those are things I pick-up and memorize about the person"*

3.2.4 Coping strategies for recognizing weak-ties. One coping strategy common among all of the participants is to ask people to introduce themselves. However, four participants mentioned that sometimes they do not feel comfortable asking for people's name. The coping strategy for these participants is to pretend to recognize people, as P1 described: *"We just get good in pretending to know who the person is. I am a professor at university. Whenever somebody on campus says hi to me, I assume this is a student, [and] I just say hi and good luck with your exams and that sort of things. Wouldn't it be nice to have more information? Yeah, sure of course!"* In situations like this, VIPs lose the chance to develop the conversation because they do not know who the other person is and what topics to talk about with her. However, sometimes VIPs continue in a conversation while pretending to know the other person, and use the conversation to look for clues to recognize him. As P5 described, *"I try to figure out by having a conversation and try to get their voice and some clue. But it is difficult because it's different from recognizing their face."*

4 FACE RECOGNITION ASSISTANT

From the formative study, we learned that VIPs consider interacting with their weak-tie connections to be very important. VIPs felt different information provided helpful clues that could help them to recognize the people whom they are interacting with. Although name and voice can help them to recognize people, they mentioned various scenarios in which that might not be the case, especially if the subject is someone they do not know well. Moreover, our participants mentioned contextual cues like the location as one of the factors that could assist them in recognizing people they know casually. However, VIPs described their weak-tie connections using more than just location, such as a person's career, how they know each other, and in general any information that is related to their previous interaction. These findings guided the design of the developed research prototype to better meet the VIPs' needs.

4.1 High-level design

The developed research prototype comprises of a wearable camera client application that runs on a smartphone worn by the user on her chest, and a server application that runs on a laptop. Because state-of-the-art computer vision algorithms are computationally expensive, all computations are done on the server in the current implementation[2]. The camera captures and sends video frames to the server, via a Wi-Fi connection, to perform the face recognition task.

The performance of a face recognition system depends significantly on the quality of the face images in the database. One way of creating the database is to populate it with pictures of the people whom the VIP wants to recognize. However, this requires that VIPs have foresight about everyone that they want the system to help them recognize. Additionally, it requires that VIPs have pictures of all of these individuals. Whereas they might have photos of strong tie connections, they may not have photos of people that they encounter infrequently and know only casually. Thus, unlike the previous works, the developed prototype creates and populates the database as VIPs interact with new people without any intervention. However, many face images extracted from videos captured in real-world scenarios are not good candidates to add to the database due to poor focus, motion blur, poor lighting, extreme head pose, and low-resolution [30,31]. In light of this, the system further processes all detected faces to filter out unacceptable (or distorted) face instances.

Whenever the user interacts with a person for the first time, the system automatically captures that individual's face, identifies undistorted clear images of her face, and adds them to the database. The database gets updated continuously in such a way that face images with higher quality for each person replace the ones already in the database. Based on the participant's feedback from the formative study on the information that would be useful for recognizing people they encounter, the system associates each person with contextual information about her interaction with the user. Thus, the system stores the interaction location, interaction time, duration and the number of interactions with each person in the database. The smartphone also records the audio data related to the user's interactions. Because participants pointed out the importance of other clues beyond this set of contextual information, VIPs can complement the system information by recording an audio description about the subject using the smartphone. Figure 1 summarizes the research prototype framework.

4.2 Implementation details

4.2.1 Detecting social interaction. To detect user involvement in a social interaction, the system detects all of the faces in the scene using MTCNN—a state-of-the-art face detection algorithm based on neural networks

[2] As phones get stronger and begin to include an AI chip [41], all these computations can be done on the phone. Moreover, researchers are trying to optimize neural networks in a way that they can fit on the mobile devices [38, 39, 40]

[27]. Next, the system estimates the distance and the head pose for each detected face. We assume that people typically would interact with the VIP at the close phase social distance (between 120 cm to 210 cm) [25] and have a head pose between -θ°_{yaw} and +θ°_{yaw}; as a result, we use 200 cm as the border for detecting the social interaction with another person. To estimate people's line of sight, we applied state-of-the-art face pose estimation based on neural networks [26]. We also used the camera model to compute other people's distance from the VIP. Additionally, the system captures contextual information (i.e., interaction time, location, and duration), and the audio snippet related to each user's interaction. The device then sends the recorded interaction information to the database.

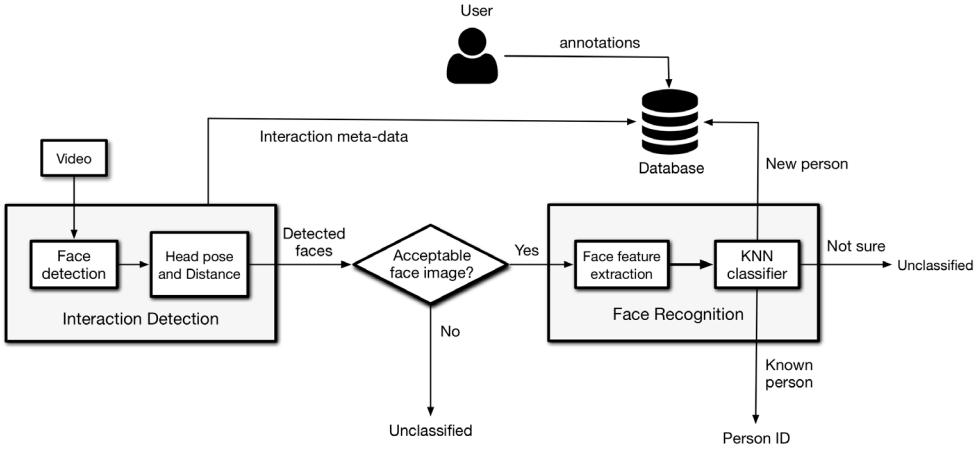

Fig. 1. The face recognition research prototype framework.

4.2.2 Populating the database without user intervention. The system will automatically store interaction information when it detects new faces. To do this without any user intervention, the system first filters away distorted face images (*i.e.,* poorly illuminated, blurry, taken with an extreme head pose, or low-resolution).

Illumination*:* Illumination variation is an issue which commonly exists in videos captured by wearable cameras [33, 34]. To address this problem, the system discards all face instances with an average brightness lower than a predefined threshold th_{lum}.

Blurriness and Poor Focus*:* Blurry face images can be the result of fast motions made by subjects in front of the camera or the camera's movement itself. To keep these face instances out of our database, we filter out all detected faces which have a low overlap with the faces in the previous frame. We calculate the Intersection over Union (IoU) as a measure of overlap between bounding boxes of face *i* in two consecutive frames:

$$IoU = \frac{R_t^i \cap R_{t+1}^i}{R_t^i \cup R_{t+1}^i} \qquad (1)$$

where R_t^i and R_{t+1}^icorrespond to the region covered by the bounding box of the subject face *i*' in frame t and $t+1$ respectively. Poor camera focus is another cause for blurry video frames. Sometimes the camera does not find the opportunity to re-focus due to quick movements made by the camera wearer. We apply the Variance of Laplacian [32] method to detect blurred face instances caused by poor camera focus. The system discards face instances with IoU lower than th_{iou}, and blurriness more than th_{blr} in the database creation.

Extreme head poses: The different head poses of people in front of the camera could lead to a considerable change in their appearance which in turn reduces the face recognition performance. To estimate people's head pose, we used a state-of-the-art head pose estimation method based on neural networks [26] which outputs the head pose in the form of three rotational angles (pitch, roll, and yaw). We apply predefined thresholds for each rotational angle to discard faces instances with extreme poses.

Resolution*:* The system also filters-out low-resolution (small in size) face images in the database creation procedure. In practice, many low-resolution faces already get removed because the system has filtered out faces that are present beyond the VIP's social distance boundary.

4.2.3 Recognizing faces. There are two testing protocols for face recognition systems 1) close-set, and 2) open-set [6]. The closed-set protocol involves having a system classify a given face to one of the face images in the database. In the open-set protocol, the given face does not necessarily exist in the database which is the case for our system. In this case, the system does not have a pre-collected database of the VIP's contacts. The system starts with no faces in the database, and as the VIP interacts with a new person who is not already in the database, the device will capture and add that person's face into the database. Under this setting, the system represents each face with discriminative large-margin features.

We used a state-of-the-art neural network [6] to extract the representative features of all the faces that pass the system's set of filters. Then, the system employs a K-nearest neighbor (KNN) classifier with a cosine similarity metric for identification. To improve the system accuracy, the result is considered to be valid whenever the classifier is confident enough, and to be discarded otherwise. Therefore, a face is considered to belong to a new person when the similarity value is below a predefined threshold th_{new}, and a face will be classified as a known face when the similarity value is above a predefined threshold (th_{known}). For values between th_{new} and th_{known}, the system refuses to classify the face instance (unrecognized). By doing this, we filter out face images that are not rejected in the previous step. In this approach, a significant number of face instances will be left unclassified. As can be seen from Fig. 2, the interaction with the VIP may be detected as two interaction intervals separated by an unrecognized interval. To avoid this, we assume that the VIP continues to interact with the same person through unrecognized intervals that are shorter than 30 seconds.

4.2.4 Annotating the interactions with additional information. Based on the VIPs' feedback from the formative study, they might need information beyond contextual clues that we have designed the system to sense to recognize the people they encounter, such as people's career, or details about their previous interactions. Therefore, the system enables users to complement the device information by recording their own description of each person in the form of a short message. To do this, users can listen to the automatically recorded information of each person, and then record and add their own desired description.

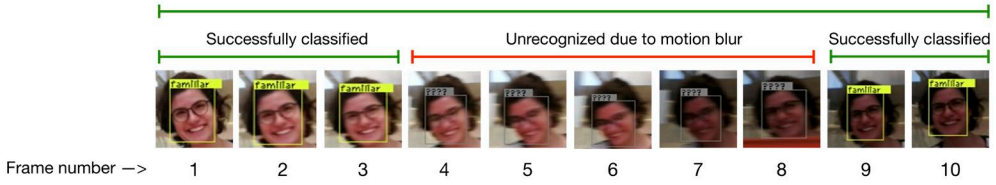

Fig. 2. The system merges detected interaction intervals that are separated by a short interval (<=30 seconds) which the system has failed to capture clear images of the subject's face. The system refuses to classify the subject's face in circumstances the subject's face is distorted due to motion blur, poor illumination, extreme head-pose, or low-resolution.

5 USER STUDY

We used the developed research prototype as a technology probe to gain an understanding about important issues involved the design and use of a face recognition assistant to help VIPs when meeting and re-

encountering new people. To this end, we designed a realistic scenario in which the visually impaired person meets new people while wearing the device. After each participant completed this scenario, we then collected feedback from them about the information that the system provides them and how the system can be improved to help them recognize others.

5.1 Participants

We recruited ten legally blind participants (5 female) with an age range of 21 to 60 (mean=35.6). Table 2 shows our participants' demographic and their vision conditions. All participants were legally blind; six of them had no vision (totally blind), two had low-vision, and the other two had ultra low-vision. We recruited the participants with the aid of Canadian National Institute for the Blind (CNIB) and BALANCE for Blind Adults, which are two volunteer agency and charitable organizations dedicated to assisting people with visual impairments. Four participants had participated in our formative study as well. Participants were compensated with $30 CAD for their involvement in our study.

Table 2. Demographics of the user study participants

Participant	Age/Sex	Occupation	Vision condition
T1	60/F	unemployed	Ultra low-vision; can see shapes only
T2	21/F	Undergraduate student	Low vision; cannot recognize faces
T3	51/M	unemployed	Blind since he was 5
T4	53/M	unemployed	Blind since he was 9
T5	51/M	Social worker	Blind since he was 10
T6	56/M	musician	Blind since 10 years ago
T7	31/F	unemployed	Low vision; cannot recognize faces
T8	38/F	Housing worker/social worker	Blind since she was 25
T9	46/F	unemployed	Ultra low-vision; can see shapes only
T10	46/M	Senior accessibility technical specialist	Blind since he was 4 years old

5.2 Study design

We designed a scenario in which participants first meet new people whom they have not interacted with previously. The scenario then presents the participants with an opportunity to re-encounter some of these individuals as well as new people. The scenario is designed in a way that simulates a user's potential encounters with acquaintances and weak tie connections. Figure. 3 demonstrates the timeline for our user study including the duration of the VIP interaction with each actor, and the time between two consecutive interactions. The scenario detail is as the following:

(1) The study facilitator (the first author) meets the participant at the lab door and walks her into the study room, where the participant meets the first actor, $A_{assistant}$, who plays the role of a researcher assistant. The facilitator introduces the participant to $A_{assistant}$ and explains that she is

there to help him conduct the study. $A_{assistant}$ greets the participant, and introduces herself (Fig 3a).

(2) The facilitator gives the participant a brief explanation of the study. At this point, the facilitator tells the participant that the goal of this study is to see if the system can help VIPs to identify people they have encountered before when they are traveling outside their home. Then, $A_{assistant}$ helps the participant to wear the camera.

(3) $A_{assistant}$ stands in front of the camera so that the system registers her face and see if the system is working properly. Then, $A_{assistant}$ reads the consent form for the participant.

(4) The participant signs the consent form. Then, the facilitator tells the participant that they would now go for a walk to the library to get a cup of coffee together and test the device.

(5) While the facilitator is packing, the second actor ($A_{distractor1}$) comes in and asks the facilitator to sign a form. The facilitator then asks $A_{distractor1}$ if they have a meeting later that day. $A_{distractor1}$ does not interact with the participants unless they initiate a conversation (Fig. 3b).

(6) The facilitator and the participant leave the study room to go get a cup of coffee together.

(7) At the reception door, they encounter the $A_{receptionist}$ actor who plays the role of the lab receptionist (Fig. 3c) and pretends to be stapling some documents. $A_{receptionist}$ says hello to the facilitator and the participant, once she sees them. The facilitator then introduces the receptionist to the participant and tells her that the participant is helping him with the study. After a short chat, the facilitator asks $A_{receptionist}$ if she wants to join them for the coffee. $A_{receptionist}$ apologizes that she cannot join them because she has to run to a class and leaves. Then, the participant wears her coat. The facilitator checks the camera again to see if it is positioned well and then they leave the lab shortly after as well.

(8) After they reach the lobby, another actor A_{silent} (Fig. 3d) holds the door for the facilitator and the participant when they are leaving the building. He does not interact with them.

(9) The facilitator and the participant cross paths with $A_{assistant}$ (Fig. 3e) on the street. Because $A_{assistant}$ is on the phone, she waives them and keeps walking.

(10) In the middle of their walk, the facilitator and the participant re-encounter $A_{receptionist}$; $A_{receptionist}$ explains to the facilitator and the participant that her classroom location got changed and she has to go back to the building (Fig. 3f).

(11) When they are near the library, the facilitator and the participant run into another actor ($A_{distractor2}$) who pretends to be the facilitator's friend. $A_{distractor2}$ greets them and after a short chat, both sides keep walking (Fig. 3g).

(12) In the library, the facilitator and the participant encounter A_{silent} again while they are in line at the coffee stand (Fig. 3h). After greetings are exchanged, A_{silent} chats with the participant about the coffee she is going to have. Because A_{silent} has already gotten his coffee, he says goodbye and leaves the coffee shop.

(13) At this point, the facilitator asks the participant to take off the camera. The facilitator stays in the line and allows the participants to take a seat. When the facilitator comes back with the coffee, the next step of the study begins.

Fig. 3. Different steps of our user study. (a) The participant meets $\mathrm{A}_{\mathrm{assistant}}$ in the study room. (b) $\mathrm{A}_{\mathrm{distractor1}}$ comes into the study room and asks the facilitator to sign a form. $\mathrm{A}_{\mathrm{distractor1}}$ does not interact with the participant unless the VIP initiates the conversation (c) The participant and the facilitator encounter the $\mathrm{A}_{\mathrm{receptionist}}$ at the reception door. (d) $\mathrm{A}_{\mathrm{silent}}$ holds the door for the facilitator and the participant when they are leaving the building. (e) The facilitator and the participant cross path with $\mathrm{A}_{\mathrm{assistant}}$. on the street (f) the facilitator and the participant re-encounter $\mathrm{A}_{\mathrm{receptionist}}$. (g) near the library, the facilitator and the participant run into $\mathrm{A}_{\mathrm{distractor2}}$ who pretends to be the facilitator's friend. (h) In the library, the facilitator and the participant encounter $\mathrm{A}_{\mathrm{silent}}$ again while they are in line at the coffee stand.

5.3 Follow-up interview

At the end of the scenario mentioned above, we conducted a semi-structured interview, where we first asked the participants to discuss every interaction they had with everyone they met during the study without the aid of the system. For actors that participants remembered their interaction with, we asked them if they remember the actor's name or not. After reviewing their interactions, participants reviewed the information that the system would have provided to them when they re-encountered $A_{assistant}$, $A_{receptionist}$, and A_{silent}. Then, we asked participants to listen to the recorded information for all remaining interactions. Participants listened to information in this order: interaction time, interaction location, interaction duration, number of interactions, and the audio snippet for that interaction. After listening to each recorded information, we asked participants to provide us with feedback about whether it would be useful in reminding them about those individuals and the encounter. Finally, we asked the participants to record any information they may want to be provided about each of the individuals they encountered during the study. The interviews took on average approximately 41 minutes. Two researchers open-coded two interview samples separately and discussed the categories. One researcher coded the remaining samples based on the agreed categories. In the next section, we present quantitative and qualitative results of our user study.

5.4 Study software

We developed an Android application as a research prototype (Fig. 4). The study software is similar to the design described in Section 4, with one major modification. We altered the implementation so that it did not interrupt the participants with information during the scenario. After the participants have completed the scenario presented in Section 5.2, we presented the participants with the information that the system could have presented to them about the persons they encountered during the walk (whereas in the actual implementation, this information would have been available during each contact). They are also given the chance to add additional information about the individuals they encountered. Thus, otherwise, the application has two modes which worked similarly to what was described in Section 4: 1) recognition, and 2) labeling. In the recognition mode, the application captures the location information in order to complement the recognition result produced by the server. In labeling mode (Fig. 4), the user can choose to listen to the information that the device has recorded for different people. By tapping on the button related to each information clue, the device plays a message using the Android TextToSpeech feature. The user can also listen to the recorded audio snippet associated with each particular interaction. Furthermore, participants can record the message that they prefer the device provides them when the subject approach them in a potential second encounter. Example audio messages are as following:

TIME: *You interacted with this person 20 minutes ago*.
LOCATION: *You interacted with this person at Bahen Centre for Information Technology*.
DURATION: *You interacted with this person for two minutes*.
FREQUENCY: *You have had two interactions with this person*.

By altering the study software in this way, we were able to use the system as a technology probe. I.e., the study software served as a tool that allowed us to ask participants to provide their reflections about the potential design and use of a face recognition assistant application, rather than an application that we evaluated how well the system performed. Additionally, the software logged video continuously during the user study to enable the frame-by-frame analysis later of each participant's interaction with the actors.

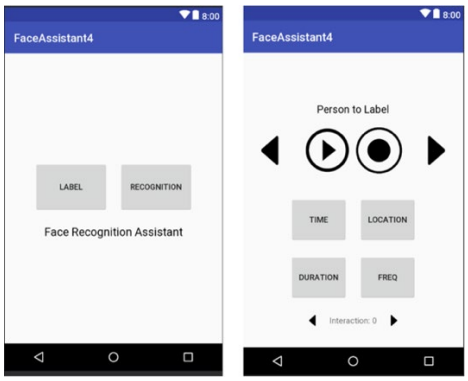

Fig. 4. The developed Android application. (Left) The application has two modes: The recognition mode and Labeling mode. In the recognition mode the device keeps track of the location information to complement the other contextual information from the server (i.e. interaction time, interaction duration, number of interaction, and the recorded audio). (Right) In the labeling mode, the user can listen to the information that the device provides and can record the information that they want to be provided when the person approaches.

5.5 Pilot study

In order to tune all of the threshold and system parameters, we first conducted two pilot studies with one sighted person and one visually impaired person (low-vision). We analyzed the recorded videos for both sessions and tuned the system variables. Table 3 shows the value for each variable.

Table 3. The system parameters set based on the pilot study

	Filtering parameters					Classifier			Database
Parameter	lu_{lth}	blr_{th}	Iou_{th}	θ_{yaw}	θ_{pitch}	K	th_{known}	th_{new}	Num of faces
Value	65	35	0.6	40°	13°	1	0.58	0.35	5

6 RESULTS

In this section, we analyze and report participant feedback about the potential design and use of a face recognition assistant. Participants provided this feedback after they completed the scripted scenario and reviewed the information that the system would have been able to provide about each actor after it registers people's faces. Thus, we analyze the system's ability to register and recognize faces and report it first to contextualize the participants' feedback. To do this, we analyzed these logged videos to evaluate the performance of the designed face recognition assistant.

6.1 Database Creation

Each participant session took on average 25 min. During each session, the number of detected face instances varied from 1718 to 4562 ($\mu = 3335$). The system rejected 73% of the detected faces and did not add them to the database or classify them. Figure 5 shows the percentage of rejected faces based on different criteria. As expected for real-life like scenarios, a large portion of the captured faces were not good candidates to be added to the database and were filtered out by the system. Note that a face image could get rejected based on more than one criteria (e.g., poor illumination, extreme pose). Figure 6a shows face instances belonging to one of the actors in the database and several discarded face instances. The system's ability to update the database also contributed to improving the database. The database updated each registered face approximately ten times with a newer face image belonging to the same person. Figure 6b demonstrates the first and the final set of face images in the database belonging to A_{silent} during T9's session. The first row of face images corresponds to the first encounter between A_{silent} and the T9 where he holds the door for her, and the second row of face images correspond to their interaction at the coffee shop. The face images from the second encounter replaced the previous ones because they were higher resolution and better illuminated. In this particular example, the database updated 21 times. The database also updated whenever a face instance with a lower head pose is available. However, due to inaccuracy in head pose estimation, this was not as effective for updating the database as updating based on illumination and resolution.

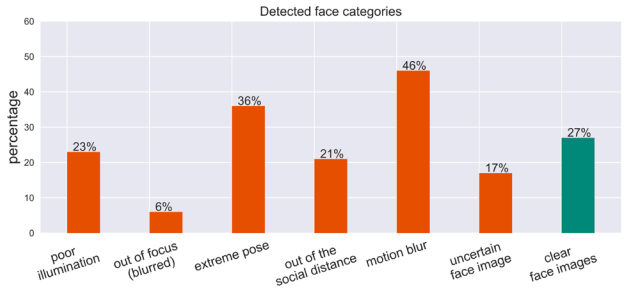

Fig. 5. Detected face categories for all sessions of the user study. From all of the detected face instances in the user study, only 27% of them (the last column) were clear enough and classified (or added to the database) by the system. Note that the sum of percentages exceeds 100% because a face instance could be categorized more than once.

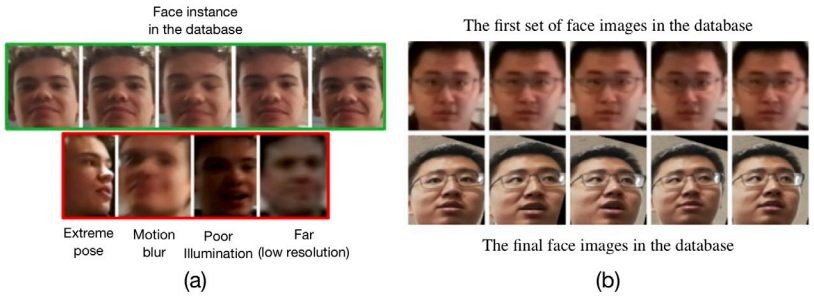

Fig. 6. (a) [First row] Face instances in the database for an actor, and [second row] discarded face instances for the same actor. (b) [First row] face instances in the database after the first encounter between the participant and $\mathbf{A_{silent}}$. and [second row] updated face instances when the participant reencountered $\mathbf{A_{silent}}$.

6.2 Registration

During our study participants interacted with five actors for the first time. Table 5 shows the registration results of all 10 user study sessions. In general, the registration performance for each interaction depends on the interaction time and how much the VIP interacted with the actor. We discuss the extent to which this result aligns with the VIPs' preference for what the system should help them to recall about each actor in subsequent sections. In this section, we first discuss the reasons for the cases that the system failed to register the subject's face.

Out of frame. Sometimes the developed prototype failed to register the subject's face in the database because it was partially or entirely out of the camera viewpoint. One reason for this issue is that the smartphone's camera usually does not have wide-angle lenses. This problem intensified in cases where the subject was closer to the participant.

Unclear face images. The system also failed to captured clear images of the actors' face even though they were within the camera viewpoint. One major reason for this issue was that the actor's face was at a wider angle than the defined thresholds with the camera. This was mostly the case because the camera was positioned at a lower height to the subject's face; thus, the subject's face had a pitch angle larger than the defined threshold to the camera. Sometimes the subject's face also exceeded the yaw threshold because the participants and the actor were not directly facing another. This was because the facilitator was also involved in the interaction or VIPs were not aware of their body position during the interaction. When the light source was located opposite the camera and illuminating the subject from behind, the subject's face appeared very dark and was discarded by the system as a result. Based on our observations, this

backlighting effect strongly changed with the camera positioning on the VIP's chest, where a slight angle towards the light source could severely darken the subject's face.

Table 5. Registration results for all of the participants' interactions. The second column shows the registration results for the participant's first encounters with actors, and the third column shows the registration results (if applicable) when the participant reencountered actors and the system failed to register the actor in the first encounter. The fourth column demonstrates the results for the cases that the system did not register the actor's face. The results for the cases in which the system registered an actor twice are shown in the fifth column.

Actors	Registered (1st encounter)	Registered (2nd encounter)	Unregistered	Over-registration
$\boldsymbol{A_{assistant}}$	10 (100%)	NA	0	1
$\boldsymbol{A_{distractor1}}$	3 (30%)	NA	7	0
$\boldsymbol{A_{receptionist}}$	8 (80%)	1	1	1
$\boldsymbol{A_{silent}}$	2 (20%)	5	3	0
$\boldsymbol{A_{distractor2}}$	6 (60%)	NA	4	0

6.3 Recognition

For our five actors, three of them ($A_{assistant}$, $A_{receptionist}$, and A_{silent}) encountered the VIP twice during the study. Table 6 shows the recognition results for the cases that the device had registered the actor during the first encounter. The system was able to recognize people who were already in the database each time they encountered of the participants again, except once for $A_{receptionist}$ in T3's user study session.

Table 6. The recognition results when the participant re-encountered actors.

Actors	Recognized / Registered
$\boldsymbol{A_{assistant}}$	10/10
$\boldsymbol{A_{recept}}$	7/8
$\boldsymbol{A_{silent}}$	2/2

6.4 Ability to recognize others through use of the system cues

At the end of the user study, we first asked the participant to discuss every interaction they had with everyone they met during the study without the aid of the system. In doing so, we were able to determine if the participants were able to recognize when they have encountered any of the actors a second time. Without the aid of the system, all participants stated that $A_{receptionist}$ is the only person whom they met twice. After reviewing their interactions, participants reviewed the information that the system would have provided them when they re-encountered $A_{assistant}$, $A_{receptionist}$, and A_{silent} for a second time (if it was available). Figure 7a shows the participants' ability to recognize the actors after listening to different information clues that the system provided. The horizontal dotted lines in Figure7 indicate the number of times that each actor was registered in the database. As can be seen from Figure 7a, after listening to all clues, the participants could identify $A_{assistant}$, $A_{receptionist}$ while they did not recognize A_{silent}. Finally, we asked participants to listen to the recorded information for all the remaining interactions and then provide

us with feedback about whether it would be useful in reminding them about those individuals and the encounter. Without the aid of the system, most participants did not remember their interactions with $A_{distractor1}$ (9 times), and $A_{distractor2}$ (7 times). Figure 7b shows the participants' ability to recognize the actors after listening to different information clues that the system provided. Note that participants listened to each clue for each actor in the same order that they appear in Figure 7 (i.e., time (T), location (L), duration (D), number of interactions (NoI), and audio (A)). The ability of participants to identify the actors might be affected by all the information accumulated by listening to the previous clues. In the following, we will discuss the helpfulness of each information clue in more detail based on the feedback collected from participants in the follow-interview.

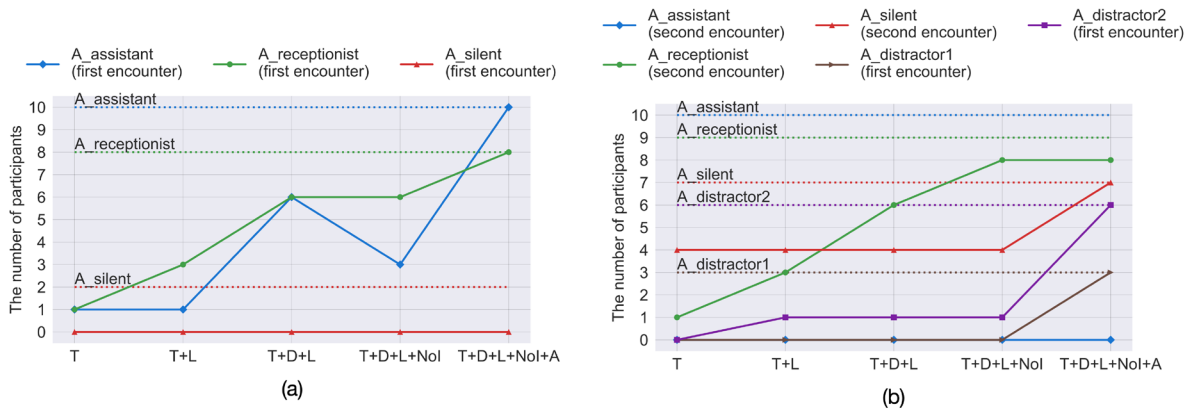

Fig. 7. (a) The participants' ability to recognize $\boldsymbol{A_{assistant}}$, $\boldsymbol{A_{receptionist}}$, and $\boldsymbol{A_{silent}}$ after listening to the information clues that pertained to their first encounter. (b) The participants' ability to recognize the actors by listening to the information clues that pertained to their second encounter with $\boldsymbol{A_{assistant}}$, $\boldsymbol{A_{receptionist}}$, and $\boldsymbol{A_{silent}}$, and their encounter with $\boldsymbol{A_{distractor1}}$, and $\boldsymbol{A_{distractor2}}$. Note that participants listened to each clue for each actor in the order indicated in x axis (i.e. time (T), location (L), duration (D), number of interactions (NoI), and audio (A)). The horizontal dotted lines indicate the number of times that each actor was registered in the database.

Interaction time: As can be seen from Figure 7b, the participants recognized their second encounter with A_{silent} four times out of seven when provided with the interaction time because it was the most recent encounter and separated in time from the others. However, participants mostly were not able to identify the encountered actors given only the interaction time. However, they still found the interaction time to be a useful clue because it helps them to limit the options for who the person could be. For example, given the time of their interaction with $A_{assistant}$, they could infer that the actor is a person whom they had met in the lab prior to the walk, excluding their interactions with $A_{distractor2}$ close to the library, $A_{receptionist}$ on the street, and A_{silent} at the coffee shop. This is particularly useful because they can conclude their guess when provided with other information like the interaction duration. Moreover, some participants were able to discover the order of interactions provided by the system and align it with what happened in their day. As T3 described: "When I listen to the one before and the one after, I place them in order then I can line that up with my memory of what I did that day. [For example,] I met this person before I went out on the walk…"

Location: The system also registered the location using the phone's GPS signal for all of the participant's interaction. Like the interaction time, participants believed that they can easily exclude the interactions outside of that location: *"It is in Bahen building so it could be either* $A_{assistant}$ *or* $A_{receptionist}$*."* Some participants (T2, T8) also believed that the interaction location is more useful than the interaction time in this manner because they always try to remember the location when they are given the time and sometimes they forget to do so. However, participants considered the location clue to be useful only when it is a named

location (such as Bahen Centre for Information Technology). The GPS failure in indoor areas is also problematic, and participants were confused when the system presented the street name instead of the building name where they met A_{silent} at the coffee shop. One participant mentioned that he needed a more detailed description of the location, such as *"Robarts library, second floor near the coffee shop."* (T4).

Interaction duration: All participants found the interaction duration to be a useful clue for interactions, because of the distinct duration in which they interacted with one person from the others. For example, participants were able to discriminate and recognize $A_{assistant}$, and $A_{receptionist}$ when they knew the corresponding duration and time of those interactions. As T3 mentioned: "The interaction lasted for two minutes, obviously I remember that I didn't have a long interaction with anyone else except her [$A_{assistant}$]".

Number of interactions: Most participants found the number of interaction confusing. The main reason for this was the fact that participants did not recognize their first interaction with A_{silent} (where she held the door), and their second interaction with $A_{assistant}$ when she waived at the participants. Thus, when the system announced two number of interactions they started to think about $A_{receptionist}$ because they thought she was the only person they encountered twice. As can be seen from Figure 7b, except for $A_{receptionist}$, the number of interactions was not helpful and contributed to three participants changing their mind after being provided by the number of interactions for $A_{assistant}$, and mistakenly thinking the information is related to $A_{receptionist}$.

Audio Snippets: After hearing the contextual information that is automatically recorded for their interactions with the actors, the participants listened to the recorded audio snippet of those interactions. All participants were able to identify all actors when listening to the audio, even the interactions that they had forgotten totally: *"Now I remember. This is the second man [in the lab]. I remember the details of the story now; he wanted you to sign something."* (T3). All participants believed the voice itself is not a factor that helped them to determine the actor's identity; in fact, it can be hard to determine someone's voice from a recording. For example, one participant (T5) mentioned that hearing the recorded voice is different than hearing that person talk live. However, participants did find that the voice was helpful in that they can infer or derive other information when listening to the conversation. For example, one participant used the audio to recognize the subject after determining the gender of the person he encountered: *"This is Jasmine ... It's a girl so I recognized it is her because I remember she is the only girl I met."* However, overall it was the conversation content that triggered their memory. *"Yeah; that was the guy I met at the coffee shop, and I remember because he asked me what flavor coffee I wanted."* (T8).

6.5 Labeling

In addition to the contextual information, the participants in our formative study mentioned several other information that they think could help them recognize the people that they interact with, such as some details about their previous interaction, and the person's profession. Thus, we asked the participants to record any information they may want to be provided about the individuals they encountered during the study. Table 7 includes what the participants recorded for $A_{assistant}$ and $A_{receptionist}$.

As can be seen from Table 7, participants found the interaction location (e.g., the lab, [our lab name], [our department name], and [our university name]) useful in describing new people they met. Two participants also used the interaction date (e.g., Nov 29) along with the interaction location. People's name is the first clue that VIPs want to be included in one's description. When the name is not available, participants would start their annotation/labeling with the subject's gender (e.g., female, the guy). Some participants also used people's work title (e.g., the research assistant), the connection point (e.g., Jennifer's friend), and some details learned during their interaction (e.g., the guy who was late for the class) to

complement the information already collected by the system. One participant described one of the actors by her attitude (e.g., the outgoing girl).

Table 7. Participants' recorded description for $A_{assistant}$ and $A_{receptionist}$ using the device labeling property.

	$A_{assistant}$	$A_{receptionist}$
T1	Jasmine	the girl in Bahen
T2	Jasmine, the research assistant in the lab	student at the lab
T3	female; met in lab, Nov 15	the guy in the lab
T4	Jasmine, a PhD student at UofT	the outgoing girl, from UofT
T5	Bahen Centre, the one who read the consent form	Jacob, met at Bahen Centre
T6	Vicky, from UofT	The guy who was late for a class
T7	This participant suffered from learning disability, and the study ended before reaching this part	
T8	Girl, who was reading me the consent form	Met outside, he was late for a class
T9	She works on the project; she read the consent form for me	Jasmine, you encountered on Nov 25
T10	Eric, Mohammad's work colleague	Jacob, Mohammad's friend

6.6 Social network management

We asked the participants to discuss whether the people registered by the system from the scenario were ones that they would have wanted to be stored in the database, in order to better understand the VIPs preference for who from their social network the system should try to recognize and the effectiveness of the developed prototype in assisting them to do so. All participants felt their interaction with $A_{assistant}$ and $A_{receptionist}$ were important enough that they would want to keep them in the database. Seven participants did not find their interaction with $A_{distractor1}$, and $A_{distractor2}$ personal enough, and preferred to remove them from the database. "When I am travelling in the subway for example; I don't think it would be useful to register people because I usually don't even talk to them; I don't want the machine to say you interacted with 20 people in the subway today." (T10) However, three participants mentioned that they would opt to include such interactions as part of their social network, because that is in their social distance, even though $A_{distractor1}$, and $A_{distractor2}$ did not interact with them directly: *"He was talking to you and I am with you. I can be with my wife, and there is a third person and the third person talks to my wife but still I am close enough by, I think proximity is a factor."* (T6) Of all of the actors, A_{silent} is the only one who did not talk when he encountered the participant for the first time (when he held the door). Almost all of our participants (9 from 10) believed that the device should only register people who communicate with them rather than everyone in their proximity. *"The people at the periphery, I don't need to know about them; ... perhaps you can put a time marker on it, so a minimum of 5 seconds interaction. (...) when you interact with someone, you*

stop, you engage, there is some kind of back and forth rather than just someone standing in front of you." (T2). However, T9 thinks there are situations that the device could help VIPs to feel more secure by registering people who are around and not talking. As she mentioned *"The user may want every single person she encounters, say to be safe; like I know a lady…she said she didn't want to use the cane in her neighborhood because she was afraid if some bad people see her condition, they might do something bad to her so maybe a lady like her wants to use a device like this to register every faces. (…) So if this device captures this person's face again and again around her, she might want to do something."*

6.7 User controls and interaction challenges

Customization is one of the parameters that contributes to the adoption of any assistive technology [34]. Based on the participants' feedback in the follow-up interview, we learned about the system aspects and parameters that VIPs want to be able to control.

6.7.1 Recognition Criteria (Whom to recognize?). We used the social distance from the user, and her line of sight as two indicators to determine if someone is interacting with the user or not. However, six participants believed that this is something they prefer to be able to adjust. Participants mentioned scenarios in which they would want the system to recognize people who are around but not necessarily interacting with (or looking at) them. As T2 described, *"Let's say I was in my building, and my sink is not working, and the maintenance guy passes me by; he is not trying to interact with me, but I know I need to talk to him; … so it would be helpful, if the machine could say, oh maintenance guy is to your left, and then I can actually initiate contact myself."* Participants also preferred to have the option to pick the recognition distance based on their needs in different situations. *"It is hard for a blind person to get the attention of a waiter when you do not know where they are. So you could possibly scan around with the device [and] maybe it could tell you where the waiter is standing roughly when they are far away."* (T6).

6.7.2 Registration Criteria (Whom to register?). The developed prototype registers people who are close enough to the user, and looking at her. However, our participants mentioned scenarios that require the system to perform based on different registration criteria. Although most participants considered proximity to be an essential factor in capturing someone in the device database, T9 as mentioned thinks of a scenario in which a VIP prefers to capture a person who is outside of the social distance and not interacting with the person to "be safe." Apart from distance and line of sight, the level at which a person engages with the VIP in an interaction is another parameter that should be adaptable across the different user's contexts. For example, T10 mentioned a subway scenario in which there are many people, but the VIP might "not even talk to them." T6 believed the presence of the person is enough even when she is not talking, whereas T2 said only people who engage with the user ("stop and engage") should be registered in the database.

6.7.3 Adaptive feedback (What clues to provide?). Participants mentioned all kinds of different information that they believed could help recognize their weak-ties. However, they might prefer a different set of cues in different contexts. In fact, three participants (T8, T9, and T10) believed that they should be able to customize the information that the device provides in different situations. As T8 described, *"I think it is important to give the person the option …. so maybe having a dropdown list of 10 items and then checking all that apply to you; there are times I am walking down the street, and I want to recall people by time and the date. Maybe next time I am traveling in the location that I know the people, I might not need to remember the time and date but maybe just the gender and the name. It depends on where you are; like here (Starbucks) I don't know anybody; if I go to the CNIB building, I know people; however, I might not necessarily remember their name; so I just want the camera only tell the name to me; not the time."*

6.7.4 Control over the information flow. Participants preferred to have control over the information flow that the device would provide them when they get involved in an interaction. For example, T10 mentioned that he does not want the devices to initiate information delivery unless he asked for it: *"I may want to put my bag go to the washroom and come back. then when I sit there, who is the room now? I am relaxed, I am calm, then I can say okay tell me what is going on."* Participants also wanted to be able to deliver the information in a gradual way. As T2 explained: *"The machine shouldn't talk all the time (…) I want to control it, you know. Like the calendar app; sometimes I only need a glance of the week [and] if I need to know more details about my schedule for a particular day then I can ask what I am going to do at 10 o'clock in the morning? I am going to a dentist; [then I can ask] okay, where is the dentist office?"*

6.7.5 The audio snippet for recognition. All participants appreciated the importance of the audio snippet in identifying the actors, and it helped them to distinguish people while they were labeling them. Yet, they did not include the audio snippet as a potential recognition clue when they were asked to list the preferred information to have when an acquaintance is approaching them. This is because of the challenges associated with interacting with recorded audio from a previous encounter to determine who the other person is. The audio recorded for the encounters that participants had with each actor in this study was on average ~41 seconds long (ranging from ~1 second to ~157 seconds); this obviously would be too long to review in full and contains irrelevant information. Furthermore, the participants would identify actors on average ~18 seconds after listening to the recorded audio. One reason for this amount of time was that sometimes the first couple of seconds of the audio snippets included the voice of someone else (the facilitator or the VIP herself) rather than the actor. Moreover, the environmental noise in real-life scenarios like in the street or coffee shop affected the participants' ability in identifying the actors faster. We also observed that the participants did not identify the subject merely by hearing her voice. In fact, participants recognized actors ~9 seconds after when the audio snippets included the subject's voice. This is still a non-trivial amount of the recording from a previous encounter that participants would need to review, and it can interfere with their current interaction with the other individual. This suggests the need to explore ways to allow the user to interact with the audio snippets efficiently without it being too disruptive. Additionally, participants considered their interaction story (e.g., "talking about coffee", "being late for the class", and "reading the consent form") as important aspects of their conversation that should be able to glean quickly from their review of the audio snippets.

6.8 System errors and the user experience

System errors and inaccuracy could lead to user frustration and eventually assistive technology abandonment [34, 35, 36]. In our system, participants were able to detect system errors in some situations, such as when it registered a non-face image in the database. Another example is when the device registered random people who did not interact with the participant (T4, T10). Although they were standing within the user's social distance, and looking at her, she might not consider that as an interaction and would want to remove them from the database. Again, participants were able to recognize this error by listening to the audio snippet. *"I have no idea what it is; at this point, I would just probably remove it"* (T10). All these observations confirm again the importance of associating the recorded audio snippet to user's interaction, which in this case would allow the users to correct the system errors in some situations.

Aside from false registration, the device might detect a person within the user's social distance but fails to register her into the database. Participants pointed out the ways that they preferred the system to act when this is the case. For example, T4, T5, and T6 preferred to know the relative location of the person, and T10 wanted to know if the other person is looking at him or not. He also pointed out that the audio could help him to review this interaction if he wants. *"If the device can record the audio I can listen to it later on and see if I can recognize it from the context"* (T10).

7 DISCUSSION

In this study, we studied how a face recognition system can help VIPs in recognizing their weak-ties, and explored the requirements and design challenges of such a system that emerge in real-life like scenarios. After learning about the users' requirements, we developed a face recognition prototype and used it as a probe to collect participant feedback about the design and use of a face recognition assistant when they meet new people. Furthermore, by conducting follow-up interviews, we learned about the system features that can complement the designed face recognition system. Unlike previous works, the developed system prototype does not require the user's effort (or a sighted person who helps the user) in order to create the system database. Moreover, whenever the user interacts with someone, the system captures different clues that might help the VIP to recognize people she interacts with in a potential second encounter.

The participants' feedback suggests that a face recognition assistant needs to be able to register and recognize people who are standing outside of the user's social distance like a scenario in which a visually impaired student attends a class and needs to communicate with the lecturer at the end of the lecture. In this case, it is crucial to obtain images of subjects' face with an acceptable resolution, because face recognition and head-pose estimation with low-resolution data is still a challenge, despite the impressive performance that achieved in recent years. Extending the distance parameter will include more people within the system's processing range, and many subjects might get added into the database even though they do not interact with the user. As a result, designers need to define more representative criteria for interaction detection. Including a talking criterion could improve the registration performance by excluding face candidates that pertain to subjects who do not talk. For example, one type of error in registration in our user study stemmed from face detection errors confusing non-face objects with a human face (e.g., an individual poster, a part of a whiteboard). Adding a talking criteria could filter out this type of error.

The environmental noise and irrelevant parts in the audio snippets led to the long recognition time. To address this problem, designers can use voice recognition to exclude irrelevant parts when someone other than the subject is talking. Furthermore, designers can leverage computer vision or audio based speaker diarization methods [47, 48] to determine who is talking in the scene and segment the audio snippets based on the speaker identity. This property is even more crucial when the VIP is involved in an interaction with more than one person. Based on the participants' feedback, the conversation content is an important element that triggers their memory. Therefore, designers should investigate efficient ways, such as the use of Natural Language Processing (NLP), to extract critical parts of the user's conversation with other people and present them to the user. It is worth noting that the interaction audio includes rich context information (e.g., background noise) that still makes it a good clue for VIPs when they are managing and labeling their social network via the face recognition assistant. However, the presentation of the extracted information from the audio snippet needs to be studied.

From the user study and the follow-up interview, we uncovered that VIPs require various user controls, such as being able to adjust recognition and registration criteria, to the set of information that the device provides in different situations. This level of variety in the user control over the device unfolds the interaction challenges associated with designing a face recognition assistance that helps VIPs in recognizing their weak-ties. Therefore, future works need to be conducted to investigate efficient ways to enable the VIP to interact and control the device based on her preference. One approach to address this is to design a context-aware system that adapts its functionality in different situations defined by the user.

Enabling the user to see what type of information the system would be able to provide them after they completed a scripted real-life scenario allowed us to collect their feedback about the design and use of the system in a contextualized manner. We were able to learn about the potential errors of a face recognition system in the wild. Our participants provided feedback that took these errors into consideration. It is

important to provide ways for users to cope with such system errors. One of the key problems that we learned about was the registration of non-faces or random people. Users should be able to remove those instances from the database. Additionally, an approach to prevent these errors in future applications is to leverage other modalities such as audio to detect the VIPs interaction with other people. Using speech detection algorithms, the face recognition assistant can add a person into the database if she actually talked to the user. Computer vision algorithms like lip motion detection also can be used to associate the detected voice with the faces available in the scene.

Although the study was carefully designed and carried out, we are aware of some limitations. We acknowledge that the current study is limited in that the study software did not provide VIPs with feedback as they interacted with actors. We framed this study in a way to gain knowledge about information clues that can assist VIPs in recognizing their weak-ties and explore the possibility of developing a system that can collect the relevant information without the user intervention. However, future studies need to also be conducted to directly evaluate real-time face recognition assistance while VIPs are involved in an interaction. Additionally, all the participants listened to different clues at the same order. This approach leads to information accumulation which makes it hard to evaluate the helpfulness of each clue in isolation.

Another limitation of the current research prototype is the fact that it relies on a server. In order to work in real-life scenarios, the research prototype is equipped with state-of-the-art computer vision algorithms which are based on neural network (i.e., face detection, face feature extraction, and head pose estimation). Using a powerful laptop for computation allowed us to test the idea of designing a face recognition system with weak-ties in mind which performs in real-time. In future work, we will investigate the possibility of using neural network optimized to fit on mobile devices [38,39,40].

Privacy concerns of people whom the user interacts with is another aspect that was not considered in this study. The face recognition system that we developed performs seamlessly—as a result, this means that surrounding people may not realize that they are being recorded. Although the actual footage is not kept on the device, passersby may still have privacy concerns. Moreover, people's face can be reconstructed from the stored face features in the database. One approach to mitigate this issue in future work is to include an indicator in the system design that lets others know about the recording. However, the effect of such an approach on VIPs and surrounding people needs more investigation.

8 CONCLUSION

In this work, we developed a face recognition research prototype that assists visually impaired people in recognizing their weak-ties. We conducted a formative study to better understand the challenges that users face with identifying their weak-ties. Based on the users' feedback from the formative study, we learned about key user requirements to guide the design of the prototype system. Using the state-of-the-art computer vision algorithms, we developed a research prototype that detects when the VIP interact with new people and populates the system database by opportunistically capturing a few clear samples of a person's face. The database also gets continuously updated to improve the face image qualities present in the database as much as possible. Moreover, the device collects contextual information (i.e., interaction time, interaction duration, interaction location), and the audio of the user's interaction to help her to recognize people she interacts with. We designed a real-life scenario in which VIPs interacted with new people and used our prototype as a probe to collect their feedback about the design and use of such an assistance when they encounter others. The VIPs' feedback provided insights about the system requirements and challenges that may arise in designing and evaluating a face recognition system that assists VIPs in recognizing a broader range of people than their close-ties.

9 REFERENCES

[1] Panchanathan, S., Chakraborty, S., & McDaniel, T. (2016). Social Interaction Assistant: A Person-Centered Approach to Enrich Social Interactions for Individuals with Visual Impairments. IEEE Journal on Selected Topics in Signal Processing, 10(5), 942–951. https://doi.org/10.1109/JSTSP.2016.2543681

[2] Wiener,W., Lawson, G.: Audition for the traveler who is visually impaired. Foundations of orientation and mobility 2 (1997) 104–169

[3] Dreer, L. E., Elliott, T. R., Fletcher, D. C., & Swanson, M. (2005). Social problem-solving abilities and psychological adjustment of persons in low vision rehabilitation. Rehabilitation Psychology, 50(3), 232–238. https://doi.org/10.1037/0090-5550.50.3.232

[4] Shinohara, K., & Wobbrock, J. O. (2011). In the shadow of misperception. In Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11 (p. 705). New York, New York, USA: ACM Press. https://doi.org/10.1145/1978942.1979044

[5] McCreadie, Claudine, and Anthea Tinker. "The acceptability of assistive technology to older people." Ageing & Society 25.1 (2005): 91-110.

[6] Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 1. 2017.

[7] Granovetter, Mark S. "The strength of weak ties." Social networks. 1977. 347-367.

[8] Morrison, Elizabeth Wolfe. "Newcomers' relationships: The role of social network ties during socialization."Academy of management Journal 45.6 (2002): 1149-1160.

[9] Haythornthwaite, Caroline. "Strong, weak, and latent ties and the impact of new media." The information society 18.5 (2002): 385-401.

[10] Granovetter, Mark S. "The Strength of Weak Ties." American Journal of Sociology, vol. 78, no. 6, 1973, pp. 1360–1380. JSTOR, JSTOR, www.jstor.org/stable/2776392.

[11] Sandstrom, G. M., & Dunn, E. W. (2014). Social interactions and well-being: The surprising power of weak ties. Personality and Social Psychology Bulletin, 40(7), 910–922. https://doi.org/10.1177/0146167214529799

[12] Crittenden, C. N., Murphy, M. L. M., & Cohen, S. (2018). Social integration and age-related decline in lung function. Health Psychology. https://doi.org/10.1037/hea0000592

[13] Dhand, A., Luke, D., Tsiaklides, M., Lang, C., & Lee, J. M. (2016, February). Patients' Social Networks Influence Timing of Hospital Arrival After Acute Ischemic Stroke: a Mixed Methods Study. In STROKE (Vol. 47). TWO COMMERCE SQ, 2001 MARKET ST, PHILADELPHIA, PA 19103 USA: LIPPINCOTT WILLIAMS & WILKINS.

[14] Zhao, Y., Wu, S., Reynolds, L., Azenkot, S., Science, I., Tech, C., & Park, M. (2018). A Face Recognition Application for People with Visual Impairments : Understanding Use Beyond the Lab, 1–14.

[15] Kramer, K. M., Hedin, D. S., & Rolkosky, D. J. (2010, August). Smartphone based face recognition tool for the blind. In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE (pp. 4538-4541). IEEE.

[16] Neto, L. B., Grijalva, F., Maike, V. R. M. L., Martini, L. C., Florencio, D., Baranauskas, M. C. C., ... Goldenstein, S. (2017). A Kinect-Based Wearable Face Recognition System to Aid Visually Impaired Users. IEEE Transactions on Human-Machine Systems, 47(1), 52–64. https://doi.org/10.1109/THMS.2016.2604367

[17] Krishna, S., Little, G., Black, J., & Panchanathan, S. (2005). A wearable face recognition system for individuals with visual impairments. In Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility - Assets '05 (p. 106). New York, New York, USA: ACM Press. https://doi.org/10.1145/1090785.1090806

[18] Chaudhry, S., & Chandra, R. (2015). Design of a mobile face recognition system for visually impaired persons. ArXiv Preprint ArXiv:1502.00756. Retrieved from http://arxiv.org/abs/1502.00756

[19] Panchanathan, S., Chakraborty, S., & McDaniel, T. (2016). Social Interaction Assistant: A Person-Centered Approach to Enrich Social Interactions for Individuals with Visual Impairments. IEEE Journal on Selected Topics in Signal Processing, 10(5), 942–951. https://doi.org/10.1109/JSTSP.2016.2543681

[20] Starner, Thad E. "Attention, memory, and wearable interfaces." IEEE pervasive computing 1, no. 4 (2002): 88-91.

[21] Shi, Weidong, Xi Wang, Xi Zhao, Varun Prakash, and Omprakash Gnawali. 2013. "Computerized-Eyewear Based Face Recognition System for Improving Social Lives of Prosopagnosics." In Proceedings of the ICTs for Improving Patients Rehabilitation Research Techniques. IEEE. doi:10.4108/icst.pervasivehealth.2013.252119.

[22] Iwamura, Masakazu, Kai Kunze, Yuya Kato, Yuzuko Utsumi, and Koichi Kise. 2014. "Haven't We Met before ? - A Realistic Memory Assistance System to Remind You of The Person in Front of You." In Augmented Human International Conference 2014, 1–4. doi:10.1145/2582051.2582083.

[23] Utsumi, Y, Y Kato, K Kunze, M Iwamura, and K Kise. 2013. "Who Are You?: A Wearable Face Recognition System to Support Human Memory." Proceedings of Augmented Human 2013, 150–53. doi:10.1145/2459236.2459262.

[24] Kurze, Martin, and Axel Roselius. 2011. "Smart Glasses Linking Real Live and Social Network's Contacts by Face Recognition." Proceedings of the 2nd Augmented Human International Conference on - AH '11, 1–2. doi:10.1145/1959826.1959857.

[25] Hall, Edward T., and E. T. Hall. "The hidden dimension, vol. 1990." NY: Anchor Books (1969).

[26] Patacchiola, Massimiliano, and Angelo Cangelosi. "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods." Pattern Recognition 71 (2017): 132-143.

[27] Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE Signal Processing Letters 23, no. 10 (2016): 1499-1503.

[28] Voykinska, Violeta, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. "How blind people interact with visual content on social networking services." In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 1584-1595. ACM, 2016.

[29] Vázquez, Marynel, and Aaron Steinfeld. "Helping visually impaired users properly aim a camera." In Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility, pp. 95-102. ACM, 2012.

[30] Hassaballah, M., and Saleh Aly. "Face recognition: challenges, achievements and future directions." IET Computer Vision 9, no. 4 (2015): 614-626.

[31] Beveridge, J. Ross, P. Jonathon Phillips, David S. Bolme, Bruce A. Draper, Geof H. Givens, Yui Man Lui, Mohammad Nayeem Teli et al. "The challenge of face recognition from digital point-and-shoot cameras." In Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on, pp. 1-8. IEEE, 2013.

[32] Pech-Pacheco, José Luis, Gabriel Cristóbal, Jesús Chamorro-Martinez, and Joaquín Fernández-Valdivia. "Diatom autofocusing in brightfield microscopy: a comparative study." In Pattern Recognition, 2000. Proceedings. 15th International Conference on, vol. 3, pp. 314-317. IEEE, 2000.

[33] Fenwick, Kent, Michael Massimi, Ronald Baecker, Sandra Black, Kevin Tonon, Cosmin Munteanu, Elizabeth Rochon, and David Ryan. "Cell phone software aiding name recall." In CHI'09 Extended Abstracts on Human Factors in Computing Systems, pp. 4279-4284. ACM, 2009.

[34] Kintsch, Anja, and Rogerio DePaula. "A framework for the adoption of assistive technology." SWAAAC 2002: Supporting learning through assistive technology (2002): 1-10.1

[35] Day, Jeffrey Jutai, William Woolrich, Graham Strong, Hy. "The stability of impact of assistive devices." Disability and rehabilitation 23, no. 9 (2001): 400-404.

[36] Phillips, Betsy, and Hongxin Zhao. "Predictors of assistive technology abandonment." Assistive technology 5, no. 1 (1993): 36-45.

[37] Kintsch, Anja, and Rogerio DePaula. "A framework for the adoption of assistive technology." SWAAAC 2002: Supporting learning through assistive technology (2002): 1-10.

[38] Wu, Jiaxiang, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. "Quantized convolutional neural networks for mobile devices." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4820-4828. 2016.

[39] Kim, Yong-Deok, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. "Compression of deep convolutional neural networks for fast and low power mobile applications." arXiv preprint arXiv:1511.06530 (2015).

[40] Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

[41] https://www.zdnet.com/article/huawei-unveils-ai-ascend-chips/

[42] Neto, Laurindo de Sousa Britto, Vanessa Regina Margareth Lima Maike, Fernando Luiz Koch, Maria Cecília Calani Baranauskas, Anderson de Rezende Rocha, and Siome Klein Goldenstein. "A Wearable Face Recognition System Built into a Smartwatch and the Visually Impaired User." In ICEIS (3), pp. 5-12. 2015.

[43] Zhang, Ning, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. "Beyond frontal faces: Improving person recognition using multiple cues." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4804-4813. 2015.

[44] Seeman, Teresa E. "Social ties and health: The benefits of social integration." Annals of epidemiology 6, no. 5 (1996): 442-451.

[45] Gurung, R., B. Sarason, and I. Sarason. "Close personal relationships and health outcomes: A key to the role of social support." Handbook of personal relationships: Theory, research and interventions (2nd ed) Chichester, UK: Wiley(1997): 547-573.

[46] OrCam. OrCam - See for Yourself. Retrieved May 10, 2019 from https://www.orcam.com/en/

[47] Bredin H, Gelly G. Improving speaker diarization of TV series using talking-face detection and clustering. InProceedings of the 24th ACM international conference on Multimedia 2016 Oct 1 (pp. 157-161). ACM.

[48] Gebru ID, Ba S, Li X, Horaud R. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. IEEE transactions on pattern analysis and machine intelligence. 2018 May 1;40(5):1086-99.