AI-Vision Towards an Improved Social Inclusion

1st Rasem Alashkar
Electrical Engineering Department
Rochester Institute of Technology
Dubai, UAE
rka4523@rit.edu

2nd Mohamed ElSabbahy
Electrical Engineering Department
Rochester Institute of Technology
Dubai, UAE
mme7875@rit.edu

3rd Ahmad Sabha
Electrical Engineering Department
Rochester Institute of Technology
Dubai, UAE
ams4413@rit.edu

4th Momen Abdelghany
Electrical Engineering Department
Rochester Institute of Technology
Dubai, UAE
mna2230@rit.edu

5th Boutheina Tlili
Electrical Engineering Department
Rochester Institute of Technology
Dubai, UAE
bktcad@rit.edu

6th Jinane Mounsef
Electrical Engineering Department
Rochester institute of Technology
Dubai, UAE
jmbcad@rit.edu

***Abstract*—The promise of Artificial Intelligence (AI)-based technologies is immense, and benefits range from efficiency profits to unprecedented improvements of quality of life in a social exclusive world, where discrimination and marginalization pose significant difficulties for building inclusive cities and providing equal access to opportunities for disabled persons. The challenges faced by visually impaired individuals in today's world are too many to count. While many tools have been introduced to help address these problems, these tools remain very costly to afford. The proposed smart glasses AI-Vision, which aim to provide the needed support in daily routine tasks utilizes machine learning algorithms in two applications of great interest. A facial expression/age/gender recognition for meetings, and a color recognition for outfit sorting are integrated in the device's framework. The generated feedback sound can be heard by the user through the bone conduction technology while avoiding disturbance of surrounding noise. The proposed AI-Vision reduces substantially the cost of a similar product while keeping an efficient recognition accuracy.**

***Index Terms*—visually impaired, smart glasses, AI-Vision, deep learning, social inclusion**

I. Introduction

Many recent studies [1]–[4] investigated the social integration in the daily life of visually impaired persons (VIP) and showed that the lack of social support is placing barriers in the way that lead to their social exclusion. A more aware society needs to be proactive in developing concrete solutions to facilitate the social inclusion of blind or low vision individuals into their ecosystem.

Today, AI holds a great promise for increasing human agency and removing a range of barriers that prevent nearly 285 million of VIP from participating equally in public life [5]. The digital industry needs to grow further in forging innovative solutions that leverage the transformative power of AI for social good. Currently, several companies are engaging in designing and developing powerful accessibility features built in to their mainstream products and services that are responsive to the needs and lifestyles of this large and growing market segment.

In this work, we propose to apply AI-based image recognition models and implement them into an assistive technology for VIP. Similar products already exist in the form of mobile applications, such as TalkDirect iCare [6], Microsoft's Seeing AI [7] and Aipoly Vision [8], and as wearable artificial vision devices, such as Orcam MyEye [9], Eyesynth [10], Horus [11] and the Sound of Vision system introduced in [12]. These products are designed to support VIP in many different scenarios, such as reading text documents, describing the user's environment and recognizing people in the vicinity of the user.

We here address a complementary scenario that is not entirely handled by current systems in the market: recognizing the age, gender and facial expression of surrounding persons for a better interpersonal communication, as well as identifying colors of objects, such as clothes, to enhance the quality of daily life. Many work have integrated facial features recognition as an assistive technology for VIP. Henrdrik et al. [13] develop a wearable sensory substitution device (SSD) that detects facial expressions of the individuals communicating with the visually impaired. This SSD arranges facial expressions into emotions, which are then converted into vibrotactile stimuli supplied from a belt that is worn on the waist. Similarly, in [14], a Social-Aware Assistant (SAA) is proposed to provide VIP with smart glasses that are equipped with a video camera and with a haptic belt for feedback to enhance their face-to-face conversations. In [15], a face recognition system is built into the Samsung Galaxy Gear smartwatch. The gear can communicate with the Galaxy Samsung Note smartphone via Bluetooth, and the resulting audio feedback can be heard through the sound outputs or through a stereo bluetooth headset. The prototype uses a library of known subjects that need to be registered prior to recognition.

On the other hand, several recent work have deployed color recognition based assistive tools for VIP. For instance, Dominguez and Graffinia design a system that can identify the predominant color of a scene, and communicate it verbally to the user through a high-end cell phone (MTD) [16]. SonarX

978-1-7281-7031-2/20/$31.00 ©2020 IEEE

[17] converts color information from images into sound. The tool converts the hue, saturation and value parameters into sound parameters that influence the perception of pitch, timbre and loudness. Finally, Abboud et al. [18] develop a novel visual to audio SSD named EyeMusic. The design uses musical notes on a pentatonic scale generated by natural instruments to convey the visual information of color and shape.

Although these tools are intended to help VIP to perceive characteristics of the environment that are usually not easily acquired without vision, they are mostly sensory-substitution devices that provide auditory or tactile representations of visual information [13]–[18]. If tactile [13], [14], these devices often generate unpleasant sensations and mostly lack color information. In the case of auditory tools [15]–[18], they require a hearing aid that connects to both ears, where only one ear can be used while the other should remain free for the perception of ambient sounds.

The main contribution of our work is in proposing a solution that combines both applications, face features recognition and color identification, in a pair of glasses that is affordable, ergonomic and easy to use. The wearable device is mounted with two Rasberry-Pi Module V2 cameras with a Sony IMX219 8-megapixel sensor that acquires a video of the user's surroundings. The video is fed wirelessly via bluetooth to a Rasberry-Pi 3 Model B microcontroller that is integrated along with the battery pack in a portable case worn on the user's waist. The Raspberry-Pi 3 applies computer vision tools in real-time while keeping a low processing power. The recognition feedback is passed to the user as a sound. When the sound arrives at the ear, it is converted into mechanical vibrations that stimulates the cochlea where it is converted into neural impulses. The neural impulses, which are in the inner ear travel along the auditory nerve to the brain where they are translated into auditory. There are two transmission pathways to transfer waves into the inner ear, either by air conduction or bone conduction [19]. We use the bone conduction Afterschokz wearable device [20], where the acoustic signal vibrates the bones of the skull to bypass the ear drums and stimulate the cochlea. Skull bone vibration can be a result of mechanical stimulation of the skull, and this is transmitted to the brain. Therefore, the user receives the audible feedback in a complete discretion without connecting the headset to the ears, but rather to the skull bones. This also prevents any interference with surrounding sounds. Finally, the proposed assistive AI-Vision is affordable with an average price of 780$ compared to similar VIP assistive products whose prices range between 2,000$ (Horus [11]) and 3,750$ (iCare [6]).

The rest of the paper is organized as follows. Section II gives an overview of the AI-Vision system. Section III introduces in details the AI-Vision design, which includes the computer vision methods, the electronic circuitry and the print design. The experimental setup and results are presented in Section IV. Finally, Section V concludes with final remarks and future recommendations.

II. AI-Vision System Overview

We developed a wearable assistive prototype for VIP with a simple interface for user interaction. The AI-Vision aims at helping visually impaired users identifying main facial features of people in their vicinity, in addition to the colors of surrounding objects. The proposed prototype consists of glasses that are attached to the Aftershockz bone conduction device [20] as shown in Fig. 1. On both sides, the glasses are mounted with two cameras that are connected to wires passing through slits on both sides of the glasses temple. The wires transfer the image signals acquired by the cameras to a controlling device that can be worn on the user's waist. The controlling device includes the microcontroller enclosed in its case and the battery pack.

The interaction flow with the smart glasses is the following: first, the user must turn on the controlling device that also activates the cameras of the glasses by pressing the power button. Once on, the user will hear an audio feedback indicating that the device is running. Then, the user presses one of two option buttons on the side of the glasses temple. To every pressed button corresponds one of the two possible applications: age/gender/expression recognition or color recognition. In either case, the user can hear an audio feedback indicating the selected option. Then, the system uses the cameras of the glasses to perceive the user's surroundings. For the first application, as soon as a face is detected, an audio feedback is given, indicating that a person's face is being framed by the camera. At this time, the user needs to stand still for a few seconds, to complete the framing. Next, the system performs the face recognition and provides an audio feedback that characterizes the age, gender and face expression of the identified person. For the second application, where the user needs to identify the color of an object, such as an outfit, the user needs to stand still for few seconds holding the object at the level of the glasses to enable the color detection system. Again, the user is notified about the color through the audio feedback. In both applications, the audio feedback is only transmitted to the user whenever a new face or object is detected by the system.

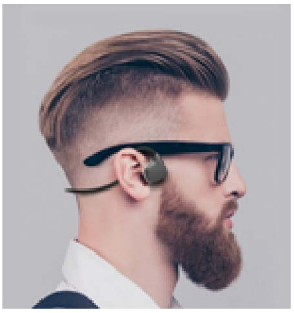

Fig. 1. AI-Vision prototype using the Aftershokz bone conduction device [20].

III. AI-Vision Design

A. Experimental Setup

For the AI-Vision's first application, three different benchmark datasets are used for training and testing the system to recognize age, gender and face expression. The FER2013 database [21] is a large-scale and unconstrained database collected automatically by the Google image search API to evaluate face expressions. All images have been registered and resized to 48*48 pixels after rejecting wrongfully labeled frames and adjusting the cropped region. FER2013 contains 28,709 training images, 3,589 validation images and 3,589 test images with seven expression labels (anger, disgust, fear, happiness, sadness, surprise and neutral). For training, the network is fed with the training dataset images in batches of 64. Each image is labeled from 0 up to 6, a value for every emotion (Fig. 2).

The Extended CohnKanade (CK+) database is the most extensively used laboratory-controlled database for evaluating face expressions [22]. CK+ contains 593 video sequences from 123 subjects. The sequences vary in duration from 10 to 60 frames and show a shift from a neutral facial expression to the peak expression. Among these videos, 327 sequences from 118 subjects are labeled with eight expression labels (neutral, anger, contempt, disgust, fear, happiness, sadness, and surprise). Using HAAR cascades, the faces in the images are detected, cropped and converted into grayscale. The resulting image sizes are 640x490 pixels. The training set is formed using 80% of the dataset, while the remaining 20% forms the test set (Fig. 2).

Finally, the Adience dataset is used to train the AI-Vision for gender and age recognition [23]. The dataset includes 26,580 photos of various subjects facing the camera in different angles (Fig. 3). All photos are downloaded from Flickr albums. Using HAAR cascades, faces are cropped and converted into grayscale. The training set contains 1,487 total images, while the test dataset includes the remaining images. The gender labels are only 2, male/female, while the age labels are 8 and are distributed over the ranges 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and 60-100.

For the AI-Vision's second application, we create the dataset needed to evaluate the color recognition. For this purpose, we include the following colors as part of the designed dataset: white, black, red, green, blue, orange, and yellow. For every color, 10 images of different shades are added totaling a number of 70 images (Fig. 4). Half of the dataset is used for training, while the other half is used for testing.

B. Proposed Methods

a) Facial Features Recognition: For the first application, a modern Convolutional Neural Network (CNN) called Mini Xception is used [24]. The Xception CNN, which is an adaptation from the already existing Inception CNN, outperforms its predecessor [25]. Here, the Inception modules are replaced with depthwise separable convolutions. This means that for every channel that exists, a pointwise 1x1 convolution is applied and for every resulting output, a 3x3 separable convolution is used alongside a rectified linear unit (ReLU) and max pooling functions. Several pooling layers within the network downsample the data provided by the convolutional layers to reduce the dimensions in the feature map, which will reduce in turn the time it takes to process the image.

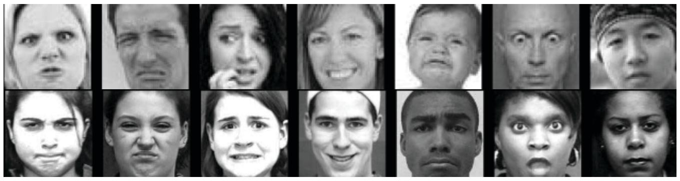

Fig. 2. Sample images from the FER2013 dataset (up) and the CK+ dataset (down). Emotions from left to right. Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral.

Fig. 3. Sample images from the Adience dataset.

We use the open-source implementation of Xception using Keras and TensorFlow, which is provided as part of the Keras Applications module2, under the MIT license [26]. The pretrained version of the network has been trained on more than a million images from the ImageNet database [27] to classify images into 1000 object categories. We train again the pre-trained Xception network on the face expression training datasets FER2013 and CK+ for 100 epochs and 20 epochs, respectively and on the age/gender training dataset Adience for 80 epochs. A 10-fold cross validation is used where data are split into 10 equal folds and the network is trained on 9 folds. It is then evaluated on the held-out fold. This is repeated 10 times so that the model can be evaluated on different held-out folds. As the size of both datasets is large, batch normalization layers are added to the Xception network to improve the performance of the trained model.

b) Color Recognition: For the color recognition application, we first compute the RGB color histogram as a feature descriptor for every image. The color histogram represents the number of pixels belonging to a fixed list of color ranges, which span the color space of the image. To classify the extracted features, we select the K-Nearest Neighbors (K-NN) algorithm, which is one of the most basic available

Fig. 4. Different shades of the colors black, blue and orange of the created color dataset.

machine learning classification methods and that do not require an extensive amount of allocated memory. This makes it an ideal choice to train our color dataset, given that the gender/age/expression application requires a larger amount of memory. The value of the hyperparameter K can be set based on the amount of training samples per label. As a default value, we use K = 3, for an optimal classification performance.

C. Sound Feedback

Information is provided to the user by directly mapping the classifiers recognition results from text into audible sounds. We have opted to use IBM Watson Text to Speech API [28]. This is a cloud service created by IBM enabling its users to convert text into a natural sounding audio. The advantages of using this API is the freedom of choosing different languages and accents, if required in the future. For the current AI-Vision version, we selected the voice AllisonV3 in US English. The resulting sound is fed through the bone conduction hearing aid, where sound signals vibrate the skull bones to produce neural impulses transmitted to the brain.

D. Hardware Design

The wiring diagram in Fig. 5 shows the different device's components and their connections.

The proposed device requires a microcontroller having at least 512MB of RAM and capable of housing more than 16GB of memory. The Raspberry-Pi 3B+ is the optimum choice to meet these requirements while being cost efficient. It holds 1GB of RAM and can support 32GB of memory.

The microcontroller is plugged into the battery pack (GeekPi Raspberry-Pi 3 UPS power supply) via its 40 main pins to be sustained with continuous uninterrupted power. The power supply contains a 1,820mAh Lipo/LiIon battery that can last from 4 to 6 hours with constant use. The battery in the power supply can be also replaced with a larger 5,000mAh or 10,000mAh battery that can last up to 24 hours. It has an on-board real time clock capable of informing the Raspberry-Pi with time even without power and an MCU chip capable of soft shutdowns, low power sleep modes and intelligent startups.

The Raspberry-Pi module communicates with the camera module (Raspberry-Pi Module V2) using the MIPI camera serial interface protocol via a ribbon cable connected to the CSI (Camera Serial Interface) port.

Finally, the Aftershokz wearable device is paired to the microcontroller via a Bluetooth wireless connection.

E. Print Design

The design of the prototype considered a casual looking model that the user can wear as modern optical glasses. The 3D print design is shown in Fig. 6. The temple has a length of 140mm and a width of 22.5mm with a curve of 44.7mm. The frame has a length of 6mm and a width of 4mm. The screen has a width of 2mm. The front of the glasses have a total length of 140mm with the upper and lower curves of 5mm and the side curves of 65mm. The curve from the lower side to the nose pads is 73.6mm and the nose pads have a length of 10mm. Behind the front screen, there are two cameras on the opposite edges with length 26mm and width 26mm. The distance between the two cameras and the center of the glasses is 50mm. The wires connecting the microcontroller to the cameras pass through two wire holes in the temple.

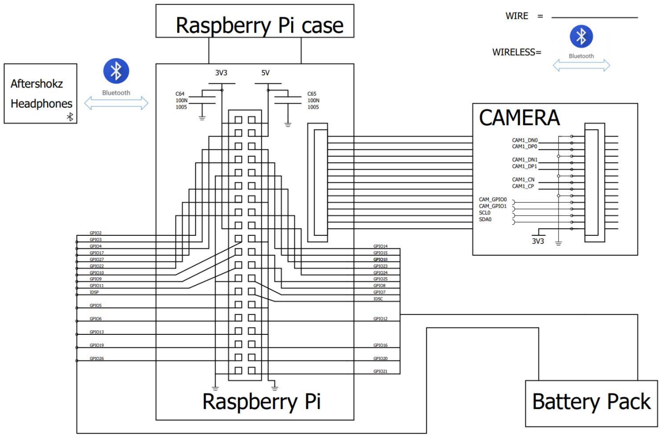

Fig. 5. Wiring diagram of the AI-Vision hardware circuit.

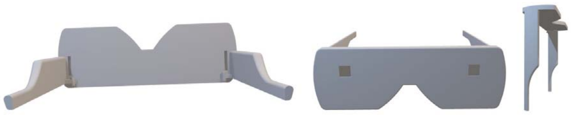

Fig. 6. 3D print design of the AI-Vision prototype.

IV. Results & Analysis

A. Facial Features Recognition

After training and testing the Mini Xception CNN model on the FER2013 and CK+ datasets, the average accuracy rates are 78.21% and 75.18%, respectively, if similar emotions are considered. Table I shows the recognition rate of every emotion for both datasets. The average accuracy rate is the highest for the FER2013 due to its larger training set compared to CK+. For the CK+ dataset, excluding the Contempt emotion increases the average recognition accuracy from 67.67% to 75.18%, which is only less by 2.62% than the FER2013 recognition rate. This is mainly due to the fact that the Contempt emotion has the smallest image sample size (18 samples) compared to the larger sample sizes of the other emotions.

The same CNN is trained and tested on the Adience dataset for age and gender recognition. The results in Table II show average recognition accuracies of 77.81% and 43.4%, respectively, for gender and age. This proves a good recognition rate for gender. On the other hand, the results of the age recognition show to be less accurate. Many factors play a role in determining the age, including hair color, moustache/beard, makeup, accessories, and lighting to name a few. This can be clearly shown by the confusion matrix that describes the

TABLE I
RECOGNITION ACCURACY OF MINI XCEPTION CNN MODEL ON FER2013 AND CK+ DATASETS.

Emotion	**Recognition Rate % (FER)**	**Recognition Rate % (CK+)**
Neutral	66.9	64.7
Happy	90.8	90.1
Angry	64.2	65.2
Sad	70.2	66.2
Disgusted	43.5	43.5
Surprised	61.3	63.3
Scared	66.4	57.9
Contempt		22.6
Average	**77.21**	**67.67**
Average w/o Contempt		**75.18**

TABLE II
AGE AND GENDER AVERAGE RECOGNITION ACCURACIES ON THE ADIENCE DATASET.

Gender **Age**	**Male**	**Female**	**Total**	**Average Recognition Rate 77.81 %**
0-2	745	682	1427	
4-6	928	1234	2162	
8-13	934	1360	2294	
15-20	734	919	1653	
25-32	2308	2589	4897	
38-43	1294	1056	2350	
48-53	392	433	825	
60-100	442	427	869	
Total	8192	9411	19487	
Average Recognition Rate 43.4 %				

TABLE III
AGE RECOGNITION CONFUSION MATRIX FORMED FROM THE ADIENCE DATASET.

	0-2	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-100**
0-2	**0.699**	0.147	0.028	0.006	0.005	0.008	0.007	0.009
4-6	0.256	**0.573**	0.166	0.023	0.010	0.011	0.010	0.005
8-13	0.027	0.223	**0.552**	0.150	0.091	0.068	0.055	0.061
15-20	0.003	0.019	0.081	**0.239**	0.106	0.055	0.049	0.028
25-32	0.006	0.029	0.138	0.510	**0.613**	0.461	0.260	0.108
38-43	0.004	0.007	0.023	0.058	0.149	**0.293**	0.339	0.268
48-53	0.002	0.001	0.004	0.007	0.017	0.055	**0.146**	0.165
60-100	0.001	0.001	0.008	0.007	0.009	0.050	0.134	**0.357**

TABLE IV
ACCURACY RATES OF THE COLOR RECOGNITION TESTED ON STILL IMAGES AND LIVE WEBCAM VIDEOS.

	White	**Black**	**Red**	**Green**	**Blue**	**Orange**	**Yellow**
Accuracy Rate %	58	69	81	78	88	32	49
Average Accuracy Rate %: 65.00							

age performance of the model on the Adience dataset (Table III). For the age ranges 15-20, 38-43, 48-53, and 60-100, the trained model does not perform well enough, which is due to the aforementioned factors.

B. Color Recognition

For the color recognition application, the K-NN model is trained and tested on the image histograms that are computed for our designed color dataset. Table IV displays the recognition rates for every color with an average accuracy of 65%. It is worth noting that the lowest recognition rates correspond to orange (32%) and yellow (49%), while the highest correspond to the RGB colors blue (88%), red (81%) and green (78%). The orange and yellow colors have low accuracies due to their similarity in shade to the red color. Moreover, many factors can affect the recognition accuracies, such as lighting, blur, and distance. Therefore, to use the color recognition application efficiently, it is recommended to maintain a good amount of light, a steady position and a close distance to the object.

V. CONCLUSION

In this paper, we presented a wearable assistive real-time system to aid the blind and low vision people with their social communication and interaction in their daily routine. The developed system consists of two distinct applications: face features and color recognitions. The first one detects a face, while the second one detects an object in the video captured by the glasses cameras. Then, the system determines the age, gender and facial expression in the first application and the color in the second application. A Mini Xception CNN and a K-NN algorithm were used to run respectively the two applications. An audio feedback is transmitted via bone conduction to notify the user of the recognized facial features or color, without interfering with ambient noises.

Several sets of experimentation were conducted on benchmark datasets to provide a preliminary evaluation of the AI-Vision applications in terms of accuracy. The system showed a satisfactory performance for most features, with average accuracy rates ranging from 65% (color recognition) to 77.81% (gender recognition). Only the age recognition had a lower accuracy rate of 43% due to many factors, including lighting and occlusions. To address this, we could use more sophisticated face detection algorithms or classifiers to improve the recognition accuracy and make the system more robust to illumination and noise conditions.

Additionally, the system was developed to show its ergonomic and economic advantages over similar products. On one hand, the user does not have to worry about using one ear to leave the other free for the perception of ambient sounds. On the other hand, the AI-Vision is cost competitive compared to other VIP assistive products that are available in the market.

Finally, we propose challenges for future work, including textual information recognition for applications, such as reading books, emails, menus, and signs but processed within the AI-Vision glasses. Furthermore, we will conduct experiments on visually impaired users to further evaluate and improve the system as an assistive device. This proposed work is inherently linked to the tenth Sustainable Development Goal (SDG), by empowering a socially excluded segment of people, unleashing opportunities for them to efficiently interact with the rest of the society, creating better conditions in which they can live, work and thrive. By utilizing and sharing the benefits of digital technology and keeping pace with advances in artificial intelligence, we can shape new frontiers for a smart and inclusive society by 2030.

ACKNOWLEDGMENT

We would like to thank Anchal Bhalla from IBM for providing help with the text to speech solution using the IBM Watson cloud service.

REFERENCES

[1] S. Sacks, L. Kekelis, and R. Gaylord-Ross. The development of social skills by blind and visually impaired students: Exploratory studies and strategies. New York, NY, USA: AFB, 1992.

[2] B. Leporini and M. Buzzi, "Home automation for an independent living: investigating the needs of visually impaired people," presented at the International Web for All, Lyon, France, Apr. 23-25, 2018.

[3] M. Pinquart and J. P. Pfeiffer, "Psychological well-being in visually impaired and unimpaired individuals: a meta-analysis," in British Journal of Visual Impairment, vol. 29, no. 1, pp.27–45, January 2011.

[4] R. L. Brown and A. E. Barrett, "Visual impairment and quality of life among older adults: an examination of explanations for the relationship," in Journals of Gerontology Series B: Psychological Sciences and Social Sciences, vol. 66, no. 3, pp. 364–373, May 2011.

[5] "Global data on visual impairment." Health World Organization. https://www.who.int/blindness/publications/globaldata/en (accessed Apr. 12, 2020).

[6] S. Panchanathan, J. Black, M. Rush, and V. Iyer, "iCare - a user centric approach to the development of assistive devices for the blind and visually impaired," presented at the IEEE International Conference on Tools with Artificial Intelligence, Sacramento, CA, USA, Nov. 5, 2003.

[7] "Seeing AI in new languages." Microsoft. https://www.microsoft.com/en-us/ai/seeing-ai (accessed Apr. 12, 2020).

[8] "Vision AI for the Blind and Visually Impaired." Aipoly. https://www.aipoly.com/ (accessed Apr. 15, 2020).

[9] J. Pauls. "An Evaluation of OrCam MyEye 2.0." American Foundation for the Blind. https://www.afb.org/aw/full-issue?id=13945 (accessed Apr. 19, 2020).

[10] "What is Eyesynth?" eyesynth. https://eyesynth.com/what-is-eyesynth/?lang=en (accessed Apr. 16, 2020).

[11] J. Beckett. "This Powerful Wearable Is a Life-Changer for the Blind." NVIDIA. https://blogs.nvidia.com/blog/2016/10/27/wearable-device-for-blind-visually-impaired (accessed Apr. 16, 2020).

[12] S. Caraiman, A. Morar, M. Owczarek, A. Burlacu, D. Rzeszotarski, N. Botezatu, P. Herghelegiu, F. Moldoveanu, P. Strumillo, and A. Moldoveanu, "Computer vision for the visually impaired: the sound of vision system," presented at the IEEE International Conference on Computer Vision, Honolulu, HI, USA, Jul. 22-25, 2017.

[13] H. P. Buimer, M. Bittner, T. Kostelijk, T. M. Van Der Geest, A. Nemri, R. J. Van Wezel, and Y. Zhao, "Conveying facial expressions to blind and visually impaired persons through a wearable vibrotactile device," in PloS one, vol. 13, no. 3, p. e0194737, March 2018.

[14] M. E. Meza-de-Luna, J. R. Terven, B. Raducanu, J. Salas, "A Social-Aware Assistant to support individuals with visual impairements during social interaction: A systematic requirements analysis," in International Journal of Human-Computer Studies, vol. 122, pp. 50-60, February 2019.

[15] L. D. Neto, V. R. Maike, F. L. Koch, M. C. Baranauskas, A. de Rezende Rocha, S. K. Goldenstein, "A wearable face recognition system built into a smartwatch and the blind and low vision users," presented at the International Conference on Enterprise Information Systems, Barcelona, Spain, Apr. 27-30, 2015.

[16] A. L. Dominguez and J. P. Graffigna, "Colors identification for blind people using cell phone," in Journal of Physics: Conference Series, vol. 332, no. 1, p. 012040, IOP Publishing, 2011.

[17] S. Cavaco, M. Mengucci, J. T. Henriques, N. Correia, and F. Medeiros, "From pixels to pitches: Unveiling the world of color for the blind," presented at the IEEE 2nd International Conference on Serious Games and Applications for Health, Vilamoura, Portugal, May 2-3, 2013.

[18] S. Abboud, S. Hanassy, S. Levy-Tzedek, S. Maidenbaum, and A. Amedi, "EyeMusic: Introducing a 'visual' colorful experience for the blind using auditory sensory substitution," in Restorative Neurology and Neuroscience, vol. 32, no. 2, pp. 247—257, January 2014.

[19] P. Henry and T. R. Letowski, "Bone conduction: Anatomy, physiology, and communication," in Army research lab aberdeen proving ground md human research and engineering directorate, No. ARL-TR-4138, 2007.

[20] Aftershokz. https://aftershokz.com/ (accessed Oct.12, 2019).

[21] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning," presented at the ACM International Conference on Multimodal Interaction, Seattle, WA, USA, Nov. 9-13, 2015.

[22] P. Lucey, J. F. Cohn, T. Kanade, and J. Saragih, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," presented at the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, Jun. 13-18, 2010.

[23] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, June 7-12, 2015.

[24] F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, Jul. 21-26, 2017.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," presented at the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, Jun. 26 - Jul. 1, 2016.

[26] B. Gupta, A. Chaube, A. Negi, and U. Goel, "Study on Object Detection using Open CV - Python," in International Journal of Computer Applications, vol. 162, no. 8, pp. 17—21, March 2017.

[27] L. Fei-Fei, J. Deng, and K. Li, "ImageNet: Constructing a large-scale image database," in Journal of vision, vol. 9, no. 8, pp. 1037–1037, August 2009.

[28] F. Santiago, P. Singh, and L. Sri. Building Cognitive Applications with IBM Watson Services: Volume 6 Speech to Text and Text to Speech. IBM Redbooks, 2017.