IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 33-40

Deep See Face Recognition – Assistive Device For Visually Impaired People

Mrs.T.Umamageswari[1], T.Deepa[2], S.Praveena[3], A.Selva Sabeena[4]

[1]Assistant Professor, Department Of Information Technology,Adhiparasakthi Engineering College
[2]Student, Department Of Information Technology, Adhiparasakthi Engineering College
[3]Student, Department Of Information Technology, Adhiparasakthi Engineering College
[4]Student, Department Of Information Technology, Adhiparasakthi Engineering College

Abstract: *We introduce the DEEP-SEE FACE framework, an assistive device designed to improve cognition, interaction and communication of visually impaired (VI) people in social encounters. The proposed approach jointly exploits computer vision algorithms (region proposal networks, ATLAS tracking and global, low level image descriptors) and deep convolutional neural networks (CNNs) in order to detect, track and recognize, in real-time, various persons existent in the video streams. The major contribution of the paper concerns a global, fixed-size face representation that takes into account various video frames while remaining independent of the length of the image sequence. To this purpose, we introduce an effective weight adaptation scheme that is able to determine the relevance assigned to each face instance, depending on the frame degree of motion/camera blur, scale variation and compression artifacts. Another relevant contribution involves a hard negative mining stage that helps us differentiating between known and unknown face identities. The experimental results, carried out on a large scale dataset, validate the proposed methodology with an average accuracy and recognition rates superior to 92%. When tested in real life, indoor/outdoor scenarios, the **DEEP-SEE FACE** prototype proves to be effective and easy to use, allowing the VI people to access visual information during social events.*

Index Terms: *Convolutional neural networks, face recognition in video streams, assistive devices for visually impaired users*

I. Introduction

Blindness is a visual impairment that affects a 0.7% of the world's population. According to the latest estimates, almost one million people in Spain suffering from visual disabilities and due to retinal diseases mentioned, about 70,000 people have total blindness. According to estimates of the World Health Organization (who), around 285 million people suffer from some sort of visual impairment, of which 39 million are blind, which means a 0.7% of the world's population? Visual impairment affects of unevenly to different age groups to be more incisive in people older than 50 years representing 65% of the total (while this group only represents 20% of the total population). The changes that occur in the vision as a result of age include:

A. Loss of the sensitivity of the retina to lighting that originates a need to use brighter lighting.
B. Opacity of the lens that causes reduced vision and annoying reflections
C. Elasticity Of the crystalline lens and loss of ability to focus
D. Degeneration of the vitreous that causes the vision of stains
E. Reduction of the capacity of the conjunctiva and lacrimal glands to adequately lubricate the eyes.

The proposed system is able to acquire information from the environment, process, interpret it and transmit acoustic messages in order to inform the VI user about the presence of a familiar face or of a known identity.

At the hardware level, the ***DEEP-SEE FACE*** system adopts our architecture initially proposed for ***DEEP-SEE*** that consists of: a mobile acquisition device (*i.e.,* a regular smartphone), a light processing unit equipped with Nvidia GPU (*i.e.,* ultra book computer) and bone conduction headphones. The proposed platform is portable, wearable and cost-effective, in order to reach the high majority of blind/visually impaired population.

In the state of the art [11], [12] various authors addressing the issue of face recognition in video streams represent a video face as a set of low level features extracted from individual frames or from the final layers of various deep neural networks [13]. Compared to still image recognition the person identification in video streams is much more challenging because of noisy frames or of unfavorable poses/viewing angles. In addition, because the same face may often include more than 100 instances, the computation time required to take them into account becomes significant. The key challenge in video face recognition is to develop a fixed-size feature representation of the face, constructed at the video level, and independent of the length of the video stream. Such a representation should allow a constant time computation in order to determine the identity of a particular individual.

The major contribution of the paper consist on an effective CNN-based weight adaptation scheme that is able to determine the relevance of the features extracted from multiple face instances, depending on the degree of motion blur, scale variations, occlusions or compression artifacts, in order to construct a compact and discriminative face representation. The proposed framework extracts per-frame video-based features using a deep face CNN model. The features are then aggregated into a global representation that can take into account the variations of the face appearance during its life cycle.

Secondly, we introduce a hard negative mining stage designed to differentiate between known faces and unknown identities. Such an issue is essential, in order to avoid false alarms, when designing a personalized learning procedure, where the users can specify their own preferences in terms of characters to be recognized.

Finally, the semantic information about the presence of a familiar is delivered with the help of acoustic warning messages, transmitted through bone conduction headphones.

The rest of the paper is organized as follows: in Section II we review the state-of-the-art approaches dedicated to the VI assistive devices based on computer vision/machine learning methods. Section III introduces the proposed architecture and describes the main steps involved: face detection, tracking, recognition and acoustic feedback. Section IV presents the experimental results obtained on a large set of videos. We show that it is possible to obtain high recognition rates on mobile wearable devices. Our system does not require any dedicated hardware architecture and can be accessible to any VI user at low cost. Finally, Section V concludes the paper and opens some directions of future work.

II. Related Work

Due to the proliferation of graphical processing units, computer vision algorithms and deep convolutional neural networks, various systems designed to increase the mobility of VI users such as ALICE [14], Mobile Vision [15] and Smart Vision [16] are based on artificial intelligence. Let us review the state-of-the-art approaches, emphasizing related strengths and limitations.

The Microsoft Kinect has been extensively used for person identification in the context of VI people. Li *et al.* [17], Cardia Neto *et al.* [18], Mian *et al.* [19], Goswami *et al.* and Berretti *et al.* [21] introduced different face recognition methods. However, such approaches are not suitable for real-time systems integrated on low processing devices.

A real-time face recognition system dedicated to blind and low-vision people is proposed in [22]. The framework integrates wearable Kinect sensors, performs face detection, and uses a temporal coherence along with a simple biometric procedure to generate a specific sound that is associated with the identified person. The underlying computer vision algorithms are tuned in order to minimize the required computational resources (memory, processing power and battery life). From this point of view, they are overcoming most state-of the-art techniques, including those proposed by Cardia Neto *et al.* [18] and Berretti *et al.* [21]. However, the range of the Kinect sensors limits the applicability of the approach to solely indoor environments.

A mobile face recognition system designed to assist the VI identification of known people is proposed in [11]. The face detection is performed using the traditional Viola-Jones algorithm with Haar-like features, while for recognition the Local Binary Patterns Histograms algorithm is used. From the experimental results it can be observed that the accuracy of the recognition module is inferior to 70% (on less than 10 classes), while the system proves to be sensitive to face poses or to different facial expressions.

Although the image -based face recognition systems have reached a high level of maturity, the methods show quickly their limitations when applied in real applications. For example, most methods prove to be highly sensitive to various changes in the illumination conditions, face poses, occlusions or low resolution. Elaborating a robust video face recognition system is still an open issue of research. Even though the deep learning methods can achieve more than 99% accuracy for face verification [27], they cannot be efficiently applied to wearable devices because of the reduced processing speed and of the significant power consumption. In the context of the ***DEEP-SEE FACE***

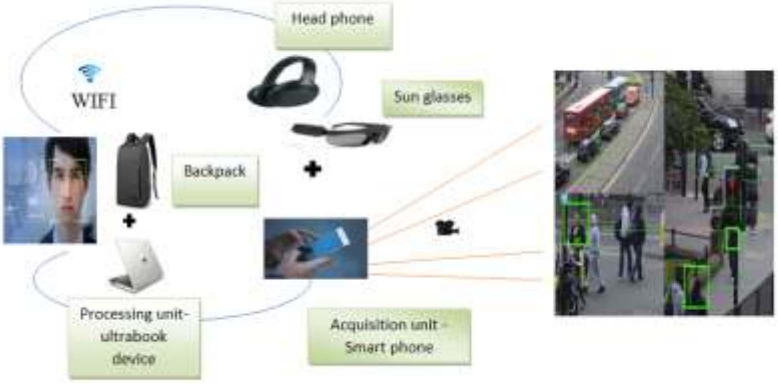
Head phone
WIFI
Sun glasses
Backpack
Processing unit-ultrabook device
Acquisition unit - Smart phone

FIGURE 1. The hardware architecture of the proposed *DEEP-SEE FACE* system

Framework, the proposed face recognition method has been specifically designed and tuned under the constraint of achieving real-time processing on portable assistive devices.

In a general manner, the state-of-the-art analysis highlights that little attention has been given to the development of a device that helps the interaction and communication of VI with other people. Moreover, the identification of faces from media, which can be highly helpful in the comprehension of the videos usually consumed by the general public, still remains a challenge for the VI community.

In this paper, we introduce the ***DEEP-SEE FACE*** framework, illustrated in Figure 1 and designed to allow VI people to access visual information during social encounters or to apprehend commonly used media.

III. PROPOSED APPROACH - DEEP-SEE FACE

Figure 2 presents the ***DEEP-SEE FACE*** architecture that involves four independent modules: face detection, multiple people tracking, people identity recognition and acoustic feedback.

A.FACE DETECTION

The face detection module is based on the Faster R-CNN with *Region Proposal Networks* (RPN) [29]. Following the default settings, we have used 3 scales (128×128, 256×256 and 512×512 pixel blocks) and 3 aspect ratios (1:1, 1:2 and 2:1) that translate to n = 9 anchors at each possible location of a face. For a feature map of size $W \times H$ (where W and H represent the width and height, respectively), we obtain a maximum number of $W \times H \times n$ proposals. As indicated in [29], the RPN training is performed using the stochastic gradient descent (SGD) for both the classification and the regression branches. We train the face detection model using the pre-trained ImageNet model of VGG [30]. The training images are resized in order to fit the GPU memory constraints based on the following scheme: *1024/max(W, H),* where W and H are the width and height of the image, respectively. The system is run for 100k iterations with a learning rate of 0.001 and for another 50k iterations at a learning rate of 0.00001.

From the results presented in Table I the following conclusions can be highlighted: (1) the lowest performance, (with a F1-score of 69.29%) is obtained by the frame-based face recognition approach. This behavior can be explained by the fact that the video stream may contain faces captured at various conditions of lighting, resolution, and pose.

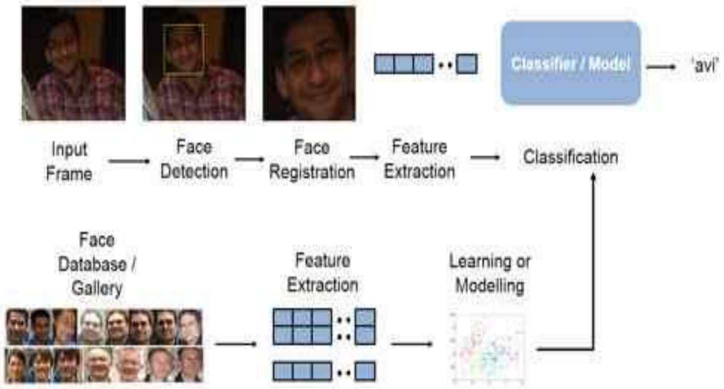

FIGURE 3. Face tracking using a modified version of the ATLAS algorithm adapted to the scenario of multiple face tracking

B. FACE TRACKING

The tracking system takes as input, at a given frame, the face bounding box indicated by the detection module (*cf.* Section III.A). Then, the goal is to determine the face position between consecutive frames. The tracking methodology is based on our previous ATLAS algorithm introduced in [31] that is adapted to work on face tracking scenarios and on multiple moving instances

We decided to use ATLAS due to its high performance and reduced computational costs. The ATLAS tracker is based on an offline-trained convolutional neural regression network that learns generic relations between various face appearances models and their associated motion patterns. The system receives as input the target and its associated search region and returns the target novel location (*i.e.*, the coordinates of the face bounding box).

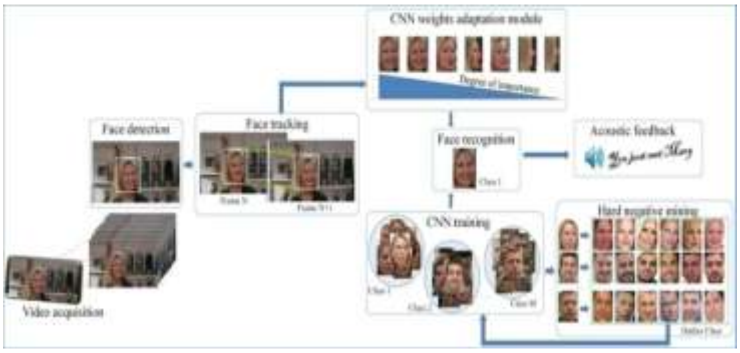

FIGURE 2. The proposed *DEEP-SEE FACE* methodological framework

The process is based on a set of comparisons between high-level features representation extracted from both faces and search regions (Figure 3). We need to emphasize that the CNN weights are modified uniquely during training (in the offline stage). In the online phase, the network weights are frozen and no fine-tuning is required. The technique is robust to important deformation, light changes or face motion and can function at more than 50fps when running on an Nvidia1050 GPU.

C. FACE RECOGNITION

Each face identified by the detection module is represented as a set of features extracted from the last layer before the classification layer of a traditional CNN. In our implementation, we have adopted the VGG16 [30] network architecture with the batch normalization strategy introduced in [32].

Let us note that other CNNs topologies can be employed. In our work, we have preferred to use a relatively standard representation, without focusing on any optimization at this stage. Instead, we have put forward the adaptation/personalization strategies. Notably, we show that such stages can be accomplished uniquely by considering the final layers of the network, with a light re-learning process.

1) WEIGHT ADAPTATION MODULE

In order to generate the set of weights, we have trained a CNN that helps us to differentiate between various face instances. We have adopted the VGG16 network architecture [30], for which we have considered only two categories, defined as ***relevant*** and ***irrelevant*** classes. They respectively correspond to high-quality frames, appropriate for recognition purposes and low-quality ones (*e.g*., blurred, profile poses…), whose impact on the recognition process should be minimized. We aim to determine for each image patch that goes through the network the probability to be assigned to the *relevant* category. Higher scores will be assigned to frontal, unblurred and unoccluded face instances.

The CNN training is performed on the Multi -Task Facial Landmark (MTFL) dataset [34] that contains 12995 face images extended with an additional 15700 faces crawled from the web. For each face in the dataset we have computed the landmark localization [35] and included in the *relevant* class only the images representing aligned faces with little variation for the yaw, roll or pitch angles (less than 25 degrees) and at a resolution superior to (128 × 128 pixels).

In order to determine the blurriness degree of the considered faces, we have adopted a non-referential sharpness (NRS) metric [36] that determines the local contrast in the neighborhood of the image edges, detected using the Sobel operator. Only faces with a NRS value inferior to 2.0 have been added to the *relevant* class. The remaining images were included in the *irrelevant* class.

In addition, both classes have been extended through a set of data augmentation techniques in order to prevent over-fitting and to enhance the generalization ability. We used the traditional data transformation methods [37] applied on image sets, such as: random cropping or horizontal flipping. For the *irrelevant* class we have adopted also the following transforms: linear motion/optical blur, face resolution (scale) variation and video compression noise in order to model the most common causes of artifacts present in video streams.

2) HARD NEGATIVE MINING

In order to deal with unknown faces, we have modified the classifier and extended the CNN output with an additional category, denoted by "*Outlier*".

Our goal is to develop a framework that is able to return the highest score for the *"Outlier"* class, against all other classes in the system, whenever the global face descriptor associated with an unknown person is applied as input.

In addition, such an approach can be useful when the detector (*cf.* Section III.A) returns false alarms. These non-face regions should be also marked as unknown instances.

In a naive approach of a weakly supervised training with stochastic gradient descent, the faces included in the *"Outlier"* set are selected from potential negative images not assigned to any category. However, it is clearly intractable to include in the unknown class all negative images from the image dataset because the categories will become unbalanced and all the new faces applied as input will be assigned to the *"Outlier"* class.

A commonly used, straightforward solution is to randomly sample the set of negative images in order to develop the unknown faces dataset. However, a limitation of this approach appears when there is a very large number of negative samples and when the known person representation is relatively good, but far from its optimum potential. In this case, most of the negative examples are considered "easy" and they will not violate the margin returning zero loss of the gradients (when performing back propagation). So, in this case, no updates of the CNN weights will be performed.

In order to deal with the above-mentioned problems we introduced a hard negative mining stage that adaptively selects the images for the *"Outlier"* class depending on the known people classes. First, using the VGG16 architecture, we perform an initial training of the CNNs for all known classes. A straightforward approach in developing the *"Outlier"* class is to apply as input to the CNNs all the face samples that are not associated to a class and to retain the first *N* hardest negative examples (*i.e.,* the images with the highest similarity score) for each category. Nevertheless, we need to take into account that an image dataset may contain multiple face instances of the same person. In the extreme case, all *N* hardest negative examples may correspond to the same face identity. In order to prevent such cases, for each category we compare, using the L2 distance, the feature vectors of the *N* negative examples in a one-to -one face verification strategy. This task eliminates duplicate instances, while allowing us to retain the faces with the highest probability of belonging to the current class.

Then, we perform again the training with this extra category. However, we have observed empirically that the CNNs will learn to solve only these particular hard cases (corresponding to the *N* negative examples when applied as input to the system) without providing significant difference from the initial network weights. We argue that mixing hard negative examples with randomly selected samples can

D. ACOUSTIC FEEDBACK

The acoustic feedback is responsible of improving the cognition of the visually impaired user about various people existent in their near surrounding. In the context of the ***DEEP-SEE*** [10] framework, the acoustic warning messages are transmitted through bone conduction headphones that satisfy the hands free and ears free conditions imposed by the VI people and enable the user to hear other external sounds from the environment.

For the ***DEEP-SEE FACE*** module, the recognized faces, are transmitted to the VI user as verbal messages, explicitly indicating the person's identity. Our major concern was to develop a warning system that is intuitive and does not require an extensive and laborious training phase. In addition, in order to provide some location information about the position of the recognized person, the warning messages are recorded in stereo using either right, left or both channels simultaneously. Thus, when the person is situated on the left (resp. right) side of the subject, the message is transmitted on the left (resp. right) channel of the bone conduction headphones. For people situated in front of the subject, the messages are transmitted in both channels.

The proposed strategy is illustrated in Figure 1, where our system transmits an acoustic warning message to the VI user in order to inform him/her about the presence of "John" within the scene.

In order not to overwhelm the VI user with redundant information, our system is designed to generate a new warning message for the same person only if the subject is present in the scene for more than 5 minutes.

III. Experimental Setup

The ***DEEP-SEE FACE*** prototype proposed in this paper shows how a robust face recognition system working directly on video streams can be used to assist the visually impaired persons when interacting with normal humans. This section highlights the major components of our system focusing our attention of the weight adaptation and the hard negative mining stages and presents the experimental results of the proposed methodology. Furthermore, tests performed in real-life scenarios, when the framework is integrated on a mobile device are presented and discussed.

A. THE BENCHMARK

Due to the novelty of the application and the unavailable free data that can be used for testing the performance of the proposed architecture, we have created a video dataset of 30 video sequences, with an average duration of 10 minutes, recorded at a resolution of 1280 x 720 pixels and with 30 fps. Five video streams have been recorded with a regular smartphone by real visually impaired users, walking in indoor/outdoor scenes, while 25 image sequences have been provided by the France national television. We need to highlight that the videos recorded in real- world conditions by the VI users are highly challenging: they are trembled, noisy, include different lighting conditions, motion blur, rotation and scale changes.

B. CNN TRAINING FOR FACE RECOGNITION

In the training phase, we have considered a dataset with 100 categories of known persons that contain faces representing user family members and friends and also some celebrities (politicians, movie stars or singers) appearing on TV. For each person, a maximum number of 800 face instances were stored in the dataset. The faces have been detected (*cf.* Section III.A) and aligned using the facial landmarks [35].

The input image size plays an important role in the training process since it can bring additional information and samples for the convolutional filters. Even though the system accuracy depends linearly on the image size, the computational resources grow quadratically. In our case, we have considered input images of size 224×224 pixels. Then, we applied batch normalization (*BN*) that solves the gradient exploding or vanishing problem and guaranties near optimal learning regime for the convolutional layers following the BN. Regarding the image batch size, this is always a tradeoff between the computational resources and the system accuracy. Experiments show [39] that keeping a constant learning rate for different min-batch sizes has a negative impact on the system's performance. Batch sizes superior to 512 or batches with single examples can lead to a significant decrease in performances. The learning rate is one of the most important hyper-parameter that needs to be adjusted when training deep neural networks, since it controls the weight variation in the direction of the gradient for a mini-batch. In our case, we used for training 50k iterations, at a learning rate of 0.0001 and a batch size of 64.

Based on transfer learning, the initialization of the CNN weights is performed using the pre-trained VGG face model that achieves state of the art results in face recognition tasks. Based on the observation of [40] that copying all but last layer of the CNN is generally the best practice for fine tuning on new small datasets, in our work we have trained only the last layer of the CNN. So, in the training stage only the weights of the final layer of the model are updated. After training, the CNN weights remain fixed.

The weight adaptation module uses for the CNN training the same parameters as the recognition module. Because, the face features are relatively compact (4096-dimensional vectors), the training process is quite efficient: training on

~1M face instances in total it takes less than 20 minutes on a GPU (Nvidia 1080Ti) mounted on a regular desktop computer.

In order to satisfy the requirements of a novel VI person using our system, the training dataset (*i.e.*, containing known people identities) can be extended/updated with additional categories, at the user's request. In this case, the CNN weights will be pre-initialized using the previously trained model. However, the training process cannot be performed by VI people or blinds and require an external effort from a technician. Once the training performed, the system can be used by the blind users without any other assistance.

C. QUANTITATIVE SYSTEM EVALUATION

The proposed face recognition system was tested on the set of 30 video streams (*cf.* Section IV.A). Because the image sequences were recorded either in crowded urban scenes or in studio with audience, more than 5.000 unknown individual were identified in the videos. In addition, the same person may appear in various environments, while in the same location various people may be present.

In the evaluation, the testing dataset is different from the face instances used for training. Initially, we have applied the face detection and tracking methods presented in Section III.A and B on the dataset of 30 videos and we cropped from each frame the regions representing faces. At this stage, we obtained 6214 faces that were tracked during the video sequence for more than one second. From the 6214 tracked faces, a number of 1108 represent known identities existent in the recognition training database. Each face instance is passed through the weight adaptation module (*cf.* Section III.C.1) in order to determine its relevance to the global face descriptor (associated to a tracked identity). Finally, the global feature vector is injected in the final layer of the CNN used in the recognition module, in order to determine the person's identity.

We have evaluated the impact of the most important parameters involved over the system's performance: the first *N* hardest negative examples used to construct the *"Outlier"* class (*cf.* Section III.C) and the *Th* $_1$ probability threshold used for assigning a face to a specific class.

Figure 5 presents the Accuracy, Recognition and F1 scores variations with respect to the various parameters involved.

Based on the results given in Figure 5 we have selected for *N* a value of 10, while the Th_1 parameter is fixed to 0.7.

In order to evaluate the influence of each components of the proposed framework on therecognition performances, we have considered for comparison:

(1) A per-frame approach that applies the face
(2) recognition algorithm to each individual frame and then takes a decision based on the dominant class;
(3) A video-based system that aggregates the face features from different instances in order to obtain a single compact representation using the baseline VGG CNN, *i.e.*, extract the *L2* normalized features followed by an average pooling [41];
(4) A compact face representation method that for each face tracked between successive frames uses a weight adaptation method as presented in Section III.C;
(5) A face recognition module that contains both the weight adaptation scheme and an *"Outlier"* class constructed with randomly selected samples.
(6) The complete framework that includes the compact face representation based on a weight adaptation scheme and constructs the *"Outlier"* category using the proposed hard negative mining methodology.

The experimental results obtained are presented in Table I.

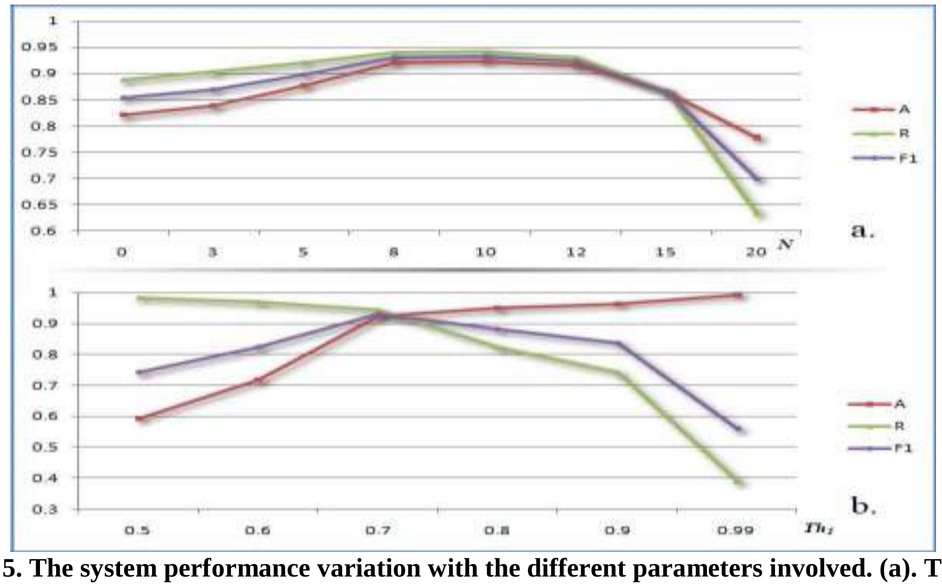

FIGURE 5. The system performance variation with the different parameters involved. (a). The first *N* hardest negative examples; (b).The probability threshold (*Th* $_1$) of assigning a face to a specific class.

IV. Conclusion

For further work and developments, we envisage to further extend the DEEP-SEE assistive device with additional functionalities that involves: inform the user when a recognized person exists the users field-of-view, navigation guidance, crossing detection or shopping assistance within large super markets. , in the recent future, to autonomously run the DEEP SEE FACE framework on a smartphone device.

References

[1]. J. Obermayer, W. Riley, O. Asif, J. Jean-Mary, “College smoking-cessation using cell phone text messaging”, J Am Coll Health. 2004;53(2):71–8.

[2]. S. Haug, C. Meyer, G. Schorr, S. Bauer, U. John, “Continuous individual support of smoking cessation using text messaging: a pilot experimental study”, Nicotine Tob Res, 11 (8) (2009), pp. 915-923.

[3]. D. Scherr, R. Zweiker, A. Kollmann, P. Kastner, G. Schreier, F.M. Fruhwald, “Mobile phone-based surveillance of cardiac patients at home”, J Telemed Telecare, 12 (5) (2006), pp. 255-261.

[4]. P. Rubel, J. Fayn, G. Nollo, D. Assanelli, B. Li, L. Restier, *et al.* “Toward personal eHealth in cardiology. Results from the EPI-MEDICS telemedicine project”,

[5]. S.C. Wangberg, E. Arsand, N. Andersson, “Diabetes education via mobile text messaging”, J Telemed Telecare, 12 (Suppl 1) (2006), pp. 55-5

[6]. P. Mohan, D. Marin, S. Sultan, A. Deen, “MediNet: personalizing the self-care process for patients with diabetes and cardiovascular disease using mobile telephony”, In: Confproc IEEE eng med biolsoc; 2008. p. 755–8.

[7]. M. Jones, J. Morris, F. Deruyter, “Mobile Healthcare and People with Disabilities: Current State and Future Needs”. Int. J. Environ. Res. Public Health.2018, 15, 515.

[8]. A World Health Organization (WHO) - Visual impairment and blindness. Available online: http://www.who.int/mediacentre/factsheets/fs282/en/ (accessed on 07.05.2018).

[9]. A. Rodríguez, J.J. Yebes, P.F. Alcantarilla, L.M. Bergasa, J. Almazán and A. Cela, “Assisting the Visually Impaired: Obstacle Detection and Warning System by Acoustic Feedback”. Sensors 2012, vol. 12, pp. 17476-17496, 10.3390/s121217476.

[10]. R. Tapu, B. Mocanu, T. Zaharia, “*DEEP-SEE*: Joint Object Detection, Tracking and Recognition with Application to Visually Impaired Navigational Assistance”,*Sensors*2017, vol. *17*, 10.3390/s17112473.

[11]. S. Chaudhry and R. Chandra, “Design of a Mobile Face Recognition System for Visually Impaired Persons”, Computer Science-Computers and Society, Computer Science - Computer Vision and Pattern Recognition, Computer Science - Human-Computer Interaction, pp. 1 – 11, 2015.

[12]. Y. Jin, J. Kim, B. Kim, R. Mallipeddi and M. Lee, “Smart cane: face recognition system for blind”. In: Proceedings of 3rd International Conference on Human-Agent Interaction, HAI 2015, 145–148. ACM, New York 2015.

[13]. Y. Rao, J. Lu and J. Zhou, “Attention-Aware Deep Reinforcement Learning for Video Face Recognition,” 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3951-3960.

[14]. R. Tapu, B. Mocanu, A. Bursuc, T. Zaharia, A Smartphone-Based Obstacle Detection and Classification System for Assisting Visually Impaired People, IEEE International Conference on Computer Vision Workshops, Sydney, 2013, pp. 444-451.

[15]. R. Manduchi, Mobile Vision as assistive technology for the blind: An experimental study. In: Proceedings of the 13th International Conference on Computers Helping People with Special Needs. 2012, pp. 9-16

[16]. J. M. H. Buf, J. Barroso, J. M. F. Rodrigues, H. Paredes, M. Farrajota, H. Fernandes, J. Jose, V. Teixeira, T. Saleiro, The smartvision navigation prototype for blind users. In JDCTA: International Journal of Digital Content Technology and its Applications, 2011, pp. 361 -372.

[17]. B. Li, A. Mian, W. Liu, and A. Krishna, “Face recognition based on Kinect,” Pattern Anal. Appl., pp. 1–11, 2015.

[18]. J.B. Cardia Neto and A. Marana, “3DLBP and HAOG fusion for face recognition utilizing Kinect as a 3D scanner,” in Proc. 30th Annu. ACM Symp. Appl. Comput., 2015, pp. 66–73.

[19]. B. Li, A. Mian, W. Liu and A. Krishna, “Using Kinect for face recognition under varying poses, expressions, illumination and disguise,” in Proc. IEEE Workshop Appl. Comput. Vision, 2013, 186–192.

[20]. G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, “On RGB-D face recognition using Kinect,” in Proc. IEEE 6th Int. Conf. Biometrics: Theory, Appl. Syst., 2013, pp. 1–6.

[21]. S. Berretti, N. Werghi, A. del Bimbo and P. Pala, “Selecting stable key-points and local descriptors for person identification using 3D face scans,”Vis. Comput., vol. 30, no. 11, pp. 1275–1292, 2014.

[22]. L. B. Neto et al., “A Kinect-Based Wearable Face Recognition System to Aid Visually Impaired Users,” in IEEE Transactions on Human-Machine Systems, vol. 47, no.1, pp52-64, 2017.

[23]. S. Chaudhry and R. Chandra, “Face detection and recognition in an unconstrained environment for mobile visual assistive system”,