# Face recognition in video streams for mobile assistive devices dedicated to visually impaired

Ruxandra Tapu[1,2], Bogdan Mocanu[1,2], Titus Zaharia[1]
[1] *ARTEMIS Department, Institute Mines - Télécom/Télécom SudParis, UMR CNRS 5157 SAMOVAR*
Evry, France
[2] *Department of Telecommunications, Faculty of ETTI, University "Politehnica" of Bucharest*
Bucharest, Romania
ruxandra.tapu@telecom-sudparis.eu, bogdan.mocanu@telecom-sudparis.eu, titus.zaharis@telecom-sudparis.eu

***Abstract*—In this paper, we introduce a novel face detection and recognition system based on deep convolutional networks, designed to improve the visually impaired users' interaction and communication in social encounters. A first feature of the proposed architecture concerns a face detection system able to identify various persons existent in the scene regardless of the subject location or pose. Then, the faces are tracked between successive frames using a CNN (*Convolutional Neural Networks*) based tracker trained offline with generic motion patterns. The system can handle face occlusion, rotation or pose variation, as well as important illumination changes. Finally, the faces are recognized, in real-time, directly from the video stream. The major contribution of the paper consists in a novel weight adaptation scheme able to determine the relevance of face instances and to create a global, fixed-size representation from all face instances tracked during the video stream. The experimental evaluation performed on a set of 30 video elements validates the approach with average detection and recognition scores superior to 85%.**

***Keywords*—face recognition in video stream, assistive device, visually impaired users, deep convolutional networks.***

## I. INTRODUCTION

According to the society of Visually Impairment and Blindness from the World Health Organization [1], more than 285 million people worldwide suffer from vision diseases, among which 39 million are completely blind. In addition, the recent statistics published by the National Eye Institute [2] in United States sustained that the number of people suffering from complete blindness has increased from 0.9 million in 2000 to 1.3 million in 2010.

The visual sense plays a significant role in guiding humans reaching the desired destination when navigating in known /unknown, indoor/outdoor environments or when interacting with other humans. For visually impaired people (VIP) the lack of vision may affect their willingness to travel and can cause emotional distress, undermining their autonomy. In this context, recognizing known human individuals is a major challenge for the VIP that prevents them to fully engage in many social activities.

With the rapid increase of powerful wearable mobile devices, complex face detection and recognition systems can be made portable. In this paper we introduce a face detection, tracking and recognition system, integrated into the ***DEEP-SEE*** [3] architecture. At the hardware level, the ***DEEP-SEE*** system is composed of a regular smartphone (*i.e.*, Samsung Galaxy S8+) that is used as acquisition device, a backpack ultrabook computer (equipped with a Nvidia 1050Ti GPU) used as a processing unit and bone conduction headphones, used to transmit the acoustic warning messages. The proposed platform is portable, wearable and cost-effective, aiming to reach the high majority of blind/visually impaired population.

At the software level, the major contribution of the paper concerns a face recognition system able to identify faces from video streams. Compared to the still image face recognition, the problem of person identification from videos is far much more challenging: noisy/blurred/occluded frames or unfavorable poses/viewing angles can seriously impact the recognition process. The framework proposed in this paper creates a fixed-size feature representation of a person, involving multiple face instances, and is independent of the face tracked/captured during the video stream. Such a representation should allow a constant time in computation. Based on an effective CNN weight adaptation scheme, the system is able to determine the relevance of a face instance, depending on the degree of motion blur, scale variations, occlusions or compression artifacts and its importance in the final compact and discriminative representation. The weighted features, extracted from various face instances, are aggregated in a global face representation used for recognition purposes.

Finally, the semantic information regarding the presence of a person within the environment is transmitted to the VIP as a set of acoustic signals.

The rest of the paper is organized as follows: in Section II we review the state-of-the-art dedicated to assistive devices that include face recognition capabilities. In Section III we describe the proposed architecture with the main steps involved: face detection, tracking, recognition and acoustic feedback. Section IV presents the experimental results obtained on a large set of videos acquired with the help of the VIP. Finally, Section V concludes the paper and opens some perspectives of future work.

## II. RELATED WORK

In the last years, due to the proliferation of computer vision algorithms and machine learning technologies, various assistive devices exploiting artificial intelligence paradigms have been proposed.

A first family of state of the art methods [4], [5], [6] is designed to perform face recognition directly on the 3D video stream acquired with the help of the Microsoft Kinect sensor. Even though the systems return satisfactory performances in the evaluation, in the context of VIP applications such approaches cannot be fully integrated in low-processing devices while fulfilling real-time capabilities.

In order to address such problems, the face recognition system introduced in [7] integrates a wearable Kinect sensor, performs face detection and uses a temporal coherence along

978-1-5386-9385-8/18/$31.00 ©2018 IEEE
DOI 10.1109/SITIS.2018.00030

with a simple biometric procedure to generate a specific sound associated with the person's identity. 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. However, the range of the Kinect sensors limits the applicability of the approach to indoor environments.

Recently, in [8], it is introduced a mobile face detection and recognition system that assists the VIP to locate and identify known persons. The face detection system is based on a set of boosted cascade classifiers with Haar-like features. The recognition process uses the Local Binary Patterns Histogram descriptor to perform subject recognition. From the experimental evaluation, we can observe that the system performance greatly depends on the size of the known person dataset. For a database with 10 classes, the system accuracy is inferior to 70% which is not acceptable in the context of VIP applications. In addition, the system is sensitive to the subject or camera motion, to face pose variation or to various facial expressions.

The emergence of deep learning techniques offers today new and promising results in the field. Thus, the approach introduced in [8] is extended in [9], where a CNN-based assistive device is proposed. First, in order to reduce the computational resources required, the video streams are resized to 283 x 500 pixels. Then, the authors propose to train a CNN architecture that incorporates three convolutional layers and two sub-sampling strategies in order to jointly detect and recognize subjects. From the experimental evaluation, it can be observed that the CNN significantly increases the detection robustness. However, the recognition system is still directly influenced by the changes in the illumination conditions or by the subject's motion. The recall and accuracy scores are inferior to 75%.

The Smart Cane face recognition system is introduced in [10]. At the hardware level, the system is composed of a video camera mounted on the subject's glasses, a computer and a vibration motor attached to the white cane. The face recognition is performed using a set of low level features obtained using a modified version of the census transform [11] and Adaboost classifier [12]. In order to inform the VIP about the presence of a known subject, the cane generates vibration patterns that are unique for each person. The system shows a high reliability in the evaluation. However, the method has never been tested with actual VIP users. Moreover, the dataset with known individuals contains only ten subjects. In addition, it requires an intensive training phase in order to become familiar with the vibration patterns.

In [13], an assistive device that performs face recognition, using a regular smartphone device, a wireless network and an audio feedback is proposed. The system is designed to recognize people situated in front of the VI user and can tolerate up to 40 degree of viewing angle (between the camera's axis and the person to be recognized) which is insufficient from the perspective of a VI user application. In addition, the smartphone needs to be hand-held, which violates the hand-free condition imposed by the blind community to any assistive device.

The FEPS (*Facial Expression Perception through Sound*) sensorial substitution system proposed in [14] extracts facial behavior expressions in a variety of social environments, in order to determine the subject's emotions and intentions. The method extracts facial landmarks that are tracked between successive frames and used to construct a 3D face model. Even though the project objectives are ambitious and the necessity of such application is obvious, the system's accuracy is relatively low, while the computational time is extensive.

A different face recognition system dedicated to VI users is introduced in [15]. First, the system performs a fast, coarse face registration by detecting eyes and nose regions. Then, the extracted regions are represented using the Local Binary Patterns, which makes it possible to construct a subject appearance model. Finally, the face identity is established based on SVM classifiers, trained using a one versus all strategy.

Although the image-based face recognition systems have reached a high level of maturity, the methods show quickly their limitations when applied in real-life 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, real-time video face recognition system is still an open issue of research. The key issue of face recognition in video streams is to build an appropriate video face representation that is able to integrate various information acquired from multiple frames together, maintaining the relevant ones while discarding the noisy information. Even, though some CNN architectures can achieve more than 99% in face verification tasks, the difficulty is to integrate such systems in light-weighted, portable assistive devices with low processing resources. Within this context, the proposed framework, described in the following section, is specifically designed in order to offer high performance on a regular wearable device.

## III. Proposed Approach

Figure 1 illustrates the proposed system architecture with the main steps involved: face detection, multiple people tracking, face recognition and acoustic feedback.

### *A. Face detection*

The face detection module is based on the Faster R-CNN [16] method with *Regional Proposal Network* (RPN) [17]. In RPN the convolution layers of a pre-trained network are followed by a *3 x 3* convolutional layer in order to perform a mapping of a receptive field (*e.g.*, 228 x 228 for VGG16 [18]) to a low dimension vector (*e.g.*, 16 for VGG16). Two *1 x 1* convolution layers are added for classification and regression branches of all spatial windows.

Following the default settings, we used 3 scales ($128^2$, $256^2$ and $512^2$ pixels 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. As indicated in [19], the RPN training is performed using the stochastic gradient descent (SGD) for both the classification and the regression branches. We initialized the VGG16 with a model pre-trained on the ImageNet [20] dataset and we have trained the CNN on the WIDER database [21]. The system is run for 100k iterations with a learning rate of 0.001.

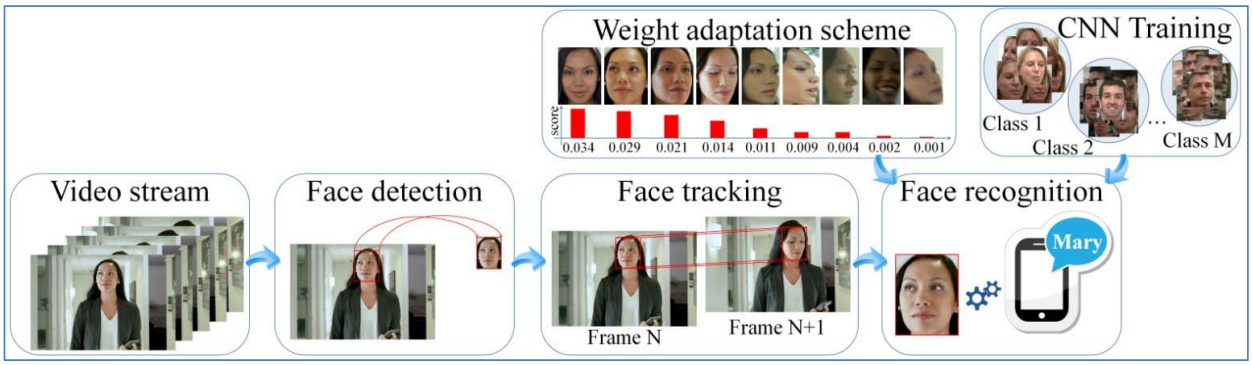


Fig. 1. The proposed system architecture with the main steps involved

### B. *Multiple faces tracking*

The proposed tracker takes as input the bounding box of the target face (indicated by the face detection module) in an initial, reference frame. The face tracking is based on our ATLAS algorithm, previously introduced in [22], that uses a regression technique, based on an offline-trained CNN that learns generic relationships between the face appearances and their associated motion patterns. In the context of an assistive device, ATLAS is adapted in order to work on face tracking scenarios and on multiple moving instances.

We decided to use ATLAS due to its high performance, obtained at the VOT2017 evaluation challenge [23], and to its reduced computational costs. In terms of processing speed, our system can track at 20 fps on a NVIDIA 1080Ti GPU.

We have considered a two-frame tracking architecture and we applied as input to the CNN, the target face as well as the search region. The network output is a set of high-level features for image representation. This output is applied as input to the fully connected layer designed to compare the features extracted from the interest face to the features from the current region of interest, in order to determine the target novel position. The network training is performed with videos and images taken from the real word, representing human faces that can undergo various transformations such as: translation, light change, occlusion or rotation.

In the online stage or test phase, the CNN weights remain unchanged and no fine-tuning is required. A tracking process based solely on motion patterns using a CNNs trained offline proves to be very fast (running at more than 50 fps), robust and accurate. However, such an approach is not able to deal with sudden camera movement, long-term occlusions or with multiple moving objects characterized by similar motion patterns and situated in the same neighboring area.

In order to overcome such limitations, we integrate rich visual and motion cues, in order to perform accurate subject position estimation [22]. The motion information is used in order to obtain an initial candidate location of the target position. Then, based on the face's previous appearance models and locations, we use a refinement strategy that adaptively modifies the object's bounding box position, size and shape, in order to avoid incorrect/false tracks.

### C. *Face recogntion*

The subject face identified by the classification module is represented as a set of high level features extracted from the final layer of the CNN architecture. We have considered the VGG16 architecture [24] with batch normalization [25] in order to leverage modern CNN network architecture with high performances.

The output of the VGG16 network is a 4096-dimensional face feature representation that is first normalized to unit vectors and fed to the aggregation module. Then, for each feature representation a weight is assigned that corresponds to its relevance to the final global face representation. Using the weight adaptation scheme, the system can take into account various face pose variations or blur, compression or motion artifacts existent in the video stream. The face recognition system is designed to determine the probability of a face to belong to a specific category.

If we denote with $X = \{x_1, x_2, \dots, x_L\}$ a face tracked during a video sequence of length $L$, where $x_k$ , $k = 1, \dots, L$ is the $k$-th frame of the video stream. For each frame $x_k$ we can determine the face normalized feature representation $f_k$ that is extracted from the VGG16 architecture.

We aim to create a global descriptor, denoted by *d(X)* and to associate it to a face that aggregates all the features extracted from multiple video frames into a compact, global face representation, defined as:

$$d(X) = \sum_k w_k \cdot f_k \,, \tag{1}$$

where $\{w_k\}_{k=1}^{L}$ is a set of weights, with $w_k$ the coefficient associated to the feature of the $k^{th}$ frame. In this way, the aggregated feature vector has the same size as a single-frame face representation extracted by the CNN.

The key element in Eq. (1) is the set of weights $\{w_k\}$. The basic solution is the naive averaging, which corresponds to assigning equal importance to each face instance such that $w_k = 1/L$. This approach is not optimal because some face poses are more representative than others. We propose a better weighting scheme that dynamically adjusts the $w_k$ parameters based on the degree of noise within the frame, face poses or viewing angles.

In order to determine the set of weights we have trained a CNN with two classes. Here again, we have adopted the VGG network architecture. The CNN assigns the face instance to two categories, denoted as ***relevant*** and ***irrelevant*** classes. During the training process, within the ***relevant*** class we have included face instances of high-quality, taken from frontal positions that are appropriate for recognition purposes. In the ***irrelevant*** class we have included blurred faces, with various motion or compression artifacts, whose

impact on the recognition process should be minimized. The output of the network is the probability of a face instance to belong to the ***relevant*** class. In this context, the higher output scores will be assigned to frontal, unblurred and unoccluded face instances.

The training of the CNN is performed of the MTFL (*Multi-Task Facial Landmark*) dataset [26] that contains 12995 face images, with various face poses for which five facial landmarks are provided. In addition, the database is extended with 15700 face images crawled from the web. In the ***relevant*** category we have included 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).

The blurriness degree of each face instance is computed using the non-referential sharpness (NRS) metric [27] that estimates the local contrast in the neighborhood of the image edges. In the ***relevant*** class we have included the face instances with a NRS score inferior to 2.0. Otherwise, the images are assigned to the ***irrelevant*** category.

In order to increase the robustness of the system, both classes have been extended with synthetically generated images, obtained with the help of some data augmentation techniques including random cropping or horizontal flipping. In addition, in order to create a higher level of generalization for the ***irrelevant*** class we have adopted also the following transforms: linear motion/optical blur, face resolution (scale) variation and video compression noise. At the end, we have obtained a database of about 1Milion images.

The weight aggregation mechanism receives as input all feature vectors extracted using the CNN architecture of the face recognition module and generates linear weights for them. If we consider the feature vector $\{f_k\}$ than the output of the weight adaptation scheme represents the instance importance $\{s_k\}$ within the global face representation. At the end, the weights are passed through a soft-max operator to generate positive, normalized weights $\{w_k\}$ with $\sum_k w_k = 1$:

$$w_k = \frac{exp(s_k)}{\sum_j exp(s_j)} \quad , \tag{2}$$

Using the proposed strategy, the number of inputs $\{f_k\}$ does not influence the size of the aggregation vector $d(X)$ which will have the same dimension as a single face feature $\{f_k\}$.

Figure 2 presents some examples of the weights computed using the proposed strategy. As it can be observed, blurred, partially occluded or profile face instances play a reduced role into the the global, aggregated face descriptor that is further used for classification purposes.

### D. Acoustic feedback

The acoustic feedback is designed to improve the VIP perception over the environment and to transmit warning messages about the recognized persons situated in the user's surroundings. We have decided to use bone conduction headphones that enable the VIP to bear the system while continuing to hear other sounds from the environment.

Verbal messages about the person identity are used, since this is the most natural manner to transmit the detected information. In addition, in order to provide hints about the location of the recognized individual, the warning messages are recorded in stereo using either right, left or both channels simultaneously. Thus, if the recognized person is situated on the left (resp. right) side of the subject than the message will be transmitted into 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.

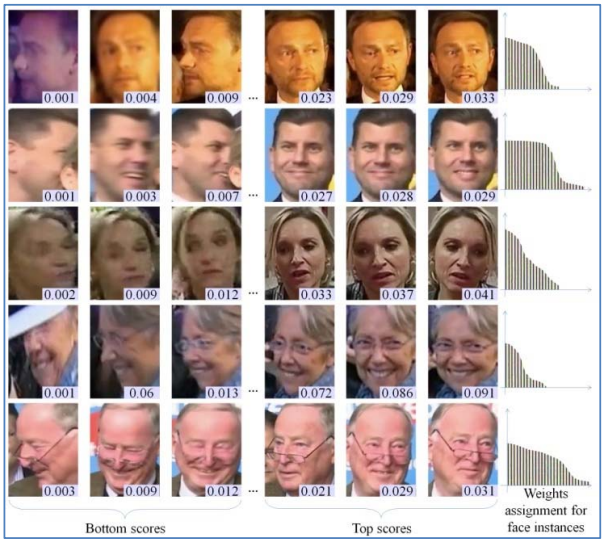


Fig. 2. Visual examples of face instances and their associated weights displayed in ascending order with respect to the video content variation

## IV. Experimental Evaluation

The proposed system has been designed to detect and recognize known individuals situated in the near surrounding of a VIP while facilitating the user interaction with other humans. The experimental evaluation is performed in real-life scenarios, with visually impaired users, when the framework is integrated on the considered wearable assistive device (***DEEP-SEE***) [3].

### A. The benckmark

Due to the novelty of the application and the unavailable free data that can be used for testing the performance of the proposed system we have developed a testing dataset with 30 video elements. The video streams were recorded by actual VIP at a resolution of 1280 x 720 pixels and with 30 fps. The average duration of an image sequence is around 10 minutes and includes in its structure multiple moving persons with various face poses. The videos are trembled, noisy, include different lighting conditions, motion blur, rotation and scale changes. Some examples of the video in the testing database are presented in Figure 3.

Fig. 3. Some examples of videos in the testing dataset

### B. CNN training in the face recognition module

The face recognition module employs the VGG16 CNN architecture. The training of the system has been performed on a set of 100 classes representing user family members and friends and also some celebrities (politicians, movie stars or singers) appearing on TV. For each person, a category was created containing a maximum number of 800 face instances. The faces included in the recognition dataset have been extracted using the face detection system (*cf.* Section III.A) and aligned based on five facial landmarks, as proposed in [26].

The face images were resized to a resolution of 224 x 224 pixels. Even though the recognition accuracy depends linearly with the image resolution, we need to make a compromise between the system accuracy and the required computational resources (*i.e.*, that grows quadratically with the image size). Then, we have applied the batch normalization (BN) to solve the gradient exploding or vanishing problem and to guaranty near optimal learning regime for the convolutional layers following the BN. The training was performed with 50k iterations, at a learning rate of 0.0001 and a batch size of 64.

Based on the transfer learning, the initialization of the weights in the CNN was performed using the pre-trained VGG face model [24] that achieves state of the art results in face recognition tasks. The CNN architecture adopted in the weight adaptation module (*i.e.,* that assigns a face to the relevant/irrelevant category) uses the same set of parameters as for the CNN used for recognition purposes.

Because the face features are relatively compact (4096-dimensional vectors), the training process is quite efficient: training on ~1M face instances in total, takes less than 20 minutes on a GPU (Nvidia 1080Ti) mounted on a regular desktop computer.

### C. Quantitative system evaluation

We have evaluated our system on the set of 30 video elements (*cf.* Section IV.A) acquired with the help of the blind society involved in our project. Because, most of the image sequence were recorded in crowded urban scenes more than 5.000 unknown individuals were identified in the videos.

The evaluation of the proposed system was performed using a set of traditional metrics such as: precision, recall and F1 score defined as follows:

$$A = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = \frac{2 \cdot A \cdot R}{A + R}, \quad (3)$$

where *TP* represent the number of true positive instances (*i.e.*, correctly recognized faces), *FP* is the number of false positive (*i.e.*, face instances incorrectly assigned to a category) and *FN* are false negative elements (*i.e.*, miss-classified faces that belong to a known class).

We start by applying the face detection and tracking methods presented in Section III A and B on the dataset of 30 video elements and we select from each video frame the face instances. We obtained 6214 faces that were tracked for more than 1 second during the video stream. From this 6214 faces a number of 1108 represent known identities existent in the training database, while the rest represent unknown persons. Then, each face instance is passed through the weight adaptation scheme (*cf.* Section III.C) in order to determine its impact on the final global descriptor associated to a tracked face. Finally, the global descriptor is inserted into the final layer of the CNN architecture in order to determine the persons' identify. The subject assignment to a category is considered as correct if the probability of belonging to that particular class is superior to 0.9.

We perform an initial experimental evaluation of each component of the proposed face recognition framework using the following strategies: (a). a *per-frame approach* in which each individual instance of a face tracked between successive frames of the video stream is considered as input to the recognition module and then a decision is taken based on the dominant category; (b). a *video-based approach* that aggregates all face instances within a global face representation and were each instance is treated with equal importance on the final global descriptor; (c). a *video-based approach with adaptive weighs assignment* that aggregates all face instances into a global descriptor and uses a weight adaptation method as presented in Section III.C.

The experimental results obtained are summarized in Table 1.

TABLE I
FACE RECOGNITION SYSTEM EVALUATION

| Condition | True Positives | False Positives | False Negatives | F1 Score |
|---|---|---|---|---|
| Frame-based method | 800 | 401 | 308 | 69.29 |
| Baseline aggregation method | 912 | 264 | 196 | 79.85 |
| Weight adaptation method | 983 | 215 | 125 | 85.25 |

From the experimental evaluation presented in Table 1, it can be observed that the proposed *video-based approach with adaptive weighs assignment* returns a F1-score superior to 85%.

Then, the 1108 known face instances were further analyzed in order to evaluate the robustness of the approach with respect to various disturbing factors. Thus, the tracked faces have been divided into the following categories: frontal face tracks, faces with important pose variation, face tracks affected by illumination changes (*e.g.*, artificial light, daylight, sunset), partially occluded faces and faces affected by important motion/camera blur. Let us underline the following observation: if a face track contains in its structure multiple disturbing factors it will be assigned to multiple categories so that the total number of faces in the ground truth database increases from 1108 instance to 1460 elements.

In Table 2 we give the TP, FP and FN parameters, while in Figure 4, we present the obtained performances on each of the considered category.

TABLE II
SYSTEM EXPERIMENTAL EVALUATION
WITH VARIOUS INDOOR / OUTDOOR CONDITIONS

| Condition | Ground Truth | True Positives | False Positives | False Negatives |
|---|---|---|---|---|
| Frontal face sequence | 518 | 502 | 14 | 16 |
| Face pose variation | 321 | 289 | 43 | 32 |
| Illumination conditions | 124 | 104 | 15 | 20 |
| Occlusion | 102 | 89 | 7 | 13 |
| Motion/camera blur | 395 | 345 | 32 | 50 |
| **TOTAL** | **1460** | **1329** | **111** | **131** |

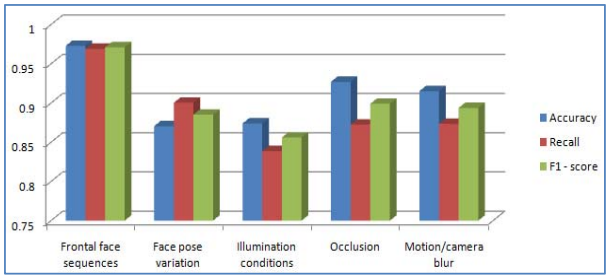


Fig. 4. The proposed system experimental evaluation

As it can be observed, our framework returns an F1 score superior to 85% regardless the lighting conditions, face pose or various types of motion existent in the scene.

## V. Conclusions and Perspectives

This paper introduces a novel CNN based face detection, tacking and recognition system designed to improve blind users' interaction and communication in social encounters. The semantic interpretation of the recognized person identity is transmitted to the VIP user through bone conducting head phones as a set of verbal acoustic messages.

From the methodological point of view, the core of the approach relies on a novel video-based face recognition approach, able to construct an effective, global and fixed-size face representation, which is independent of the length of the image sequence. A weight adaptation scheme is proposed, able to adaptively assign a weight to each face instance depending on the video content variation.

The experimental evaluation performed on a large set of video elements acquired with the help of visually impaired users validate the approach that returns an average accuracy and recall scores superior to 85% regardless the lighting conditions, face pose or various types of motion existent in the scene.

For further work and development we envisage extending the framework with additional capabilities such as: shopping or navigation assistance, crossing detection or improving the GPS navigation accuracy using computer vision algorithms.

## References

[1] World Health Organization.: Towards universal eye health: a global action plan, 2014 to 2019 report.

[2] United States National Eye Institute: United States prevalent cases of blindness (in thousands): Changes of cases between 2000 and 2010, 2016.

[3] 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.*

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

[5] 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.

[6] 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, pp. 186–192.

[7] 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.

[8] 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.

[9] S. Chaudhry and R. Chandra, "Face detection and recognition in an unconstrained environment for mobile visual assistive system", Applied Soft Computing, vol. 53, pp. 168-180, 2017, 10.1016/j.asoc.2016.12.035

[10] 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, pp. 145–148. ACM, New York 2015.

[11] R. Zabih, J. Woodfill, "Non-Parametric Local Transforms for Computing Visual Correspondence", Proc. Third European Conf. Computer Vision, pp. 150-158, 1994.

[12] J. Zhu, S. Rosset, H. Zou, T. Hastie, "Multi-Class Adaboost", 2005.

[13] K. M. Kramer, D. S. Hedin and D. J. Rolkosky, "Smartphone based face recognition tool for the blind," 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, pp. 4538-4541, 2010.

[14] M.I. Tanveer, A.I. Anam, A.M. Rahman, S. Ghosh, M. Yeasin: "Feps: A sensory substitution system for the blind to perceive facial expressions". In Proceedings of the 14th International ACM Conference on Computers and Accessibility, pp. 207–208, New York, NY, USA, 2012.

[15] G. Fusco, N. Noceti and F. Odone, "Combining Retrieval and Classification for Real-Time Face Recognition," 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, Beijing, pp. 276-281, 2012.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in NIPS, 2015.

[17] H Jiang and E. G. Learned-Miller, "Face detection with the faster R-CNN," 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, pp. 650-657, 2017, 10.1109/FG.2017.82.

[18] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In ICLR, 2015.

[19] H Jiang and E. G. Learned-Miller, "Face detection with the faster R-CNN," 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, pp. 650-657, 2017, 10.1109/FG.2017.82.

[20] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.

[21] S. Yang, P. Luo, C. C. Loy and X. Tang, "WIDER FACE: A Face Detection Benchmark," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, 2016, pp. 5525-5533.

[22] B. Mocanu, R. Tapu and T. Zaharia, "Single object tracking using offline trained deep regression networks," *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA),* Montreal, QC, 2017, pp. 1-6.

[23] M. Kristan *et al.*, "The Visual Object Tracking VOT2017 Challenge Results," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW),* Venice, 2017, pp. 1949-1972.

[24] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In ICLR, 2015.

[25] S. Ioffe and C. Szegedy,"Batch normalization: Accelerating deep network training by reducing internal covariate shift", Proceedings of the 32nd International Conference on International Conference on Machine Learning, vol. 37, pp. 448-456, 2015.

[26] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. "Facial landmark detection by deep multi-task learning". In ECCV, pp. 94–108, 2014.

[27] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB), "IEEE Trans. Image Process, vol. 18, no. 4, pp. 717–728, 2009.