IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2018, accepted September 11, 2018, date of publication September 17, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2870334

DEEP-SEE FACE: A Mobile Face Recognition System Dedicated to Visually Impaired People

BOGDAN MOCANU[1,2], (Member, IEEE), RUXANDRA TAPU[1,2], (Member, IEEE), AND TITUS ZAHARIA[1], (Member, IEEE)

[1]ARTEMIS Department, Institut Mines-Télécom/Télécom SudParis, UMR CNRS 5157 SAMOVAR, 91000 Évry, France
[2]Telecommunication Department, Faculty of ETTI, University Politehnica of Bucharest, 060042 Bucharest, Romania

Corresponding author: Ruxandra Tapu (ruxandra.tapu@telecom-sudparis.eu)

This work was supported by the University Politehnica of Bucharest, through the Excellence Research Grants Program, UPB—GEX identifier: UPB–EXCELENȚĂ–2017 under Grant 40/25.09.2017.

ABSTRACT In this paper, 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, and low-level image descriptors) and deep convolutional neural networks 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 the account of 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 data set, 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

In recent years, several frameworks based on mobile platforms and dedicated to healthcare services have emerged. The novel technologies developed aim at reducing the costs of the health sector, by increasing the empowerment of people and, in the same time, by improving the monitoring of patients with chronic diseases. Through the continuous assessment of symptoms, such systems can help the patients to managing their condition by their own, without needing direct supervision of specialized healthcare personnel.

Currently, the patient monitoring systems based on internet of things (IoT) or cyber physical systems (CPS) are attracting considerable attention from the scientific community. Such emerging technologies have been used to various purposes: facilitate smoking cessation [1], [2], monitor patients with chronic heart failure [3], detect early signs of arrhythmia or ischemia [4], provide diabetes education [5] or monitor relevant physiological markers [6].

With a few notable exceptions (mental health and autism), the people with disabilities have not been the primary target of the emerging mobile health applications. However, individuals with disabilities are likely to engage in behaviors that can put their health at risk [7] and there is a strong need of technologies that can improve their daily-life conditions, enable social relations, and increase their degree of autonomy and safety.

In this paper, we focus on a particular case of disability, which is the visual impairment. Nowadays, more than 285 million people worldwide suffer from visual impairment (VI) [8] with 39 million of blinds and 246 million people with low vision. The World Health Organization estimates that by the year of 2020 the number of individuals affected by VI will significantly increase [9]. The visually impaired people adapt to normal life by using traditional assistive aids, such as white canes or walking dogs. The white cane is preferred because it is easy to use, cheap and widely accepted

by the blind community. However, such an assistive element shows quickly its limitations when confronted with the high diversity of situations that can occur in current urban scenes. Moreover, the white cane cannot provide additional information to users such as the degree of danger of the encountered obstacles or recognition of persons that are present in the scene. In the absence of such information, the VI always travels on known paths while trying to guess the identity of the persons encountered. When a VI user arrives in a social setting, the conversation has to be interrupted in order to announce which people are present.

In this paper, we introduce ***DEEP-SEE FACE***, a novel assistive device based on computer vision algorithms and offline-trained deep convolutional neural network that extends the previously proposed ***DEEP-SEE*** [10] architecture with a face recognition module. ***DEEP-SEE FACE*** is able to identify in real-time, from video streams, a set of characters, which can be pre-defined by the user and which may correspond to either familiar people that the VI user may encounter in real life or to celebrities appearing in media streams. So, two challenges are addressed: (1) detect and recognize familiar people when navigating in indoor or outdoor environment; (2) acquire additional information about the identity of various people/celebrities appearing on the media broadcasted at TV or over internet.

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*** [10] 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 and Marana [18], Li *et al.* [19], Goswami *et al.* [20] 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 and Marana [18] and Berretti *et al.* [21].

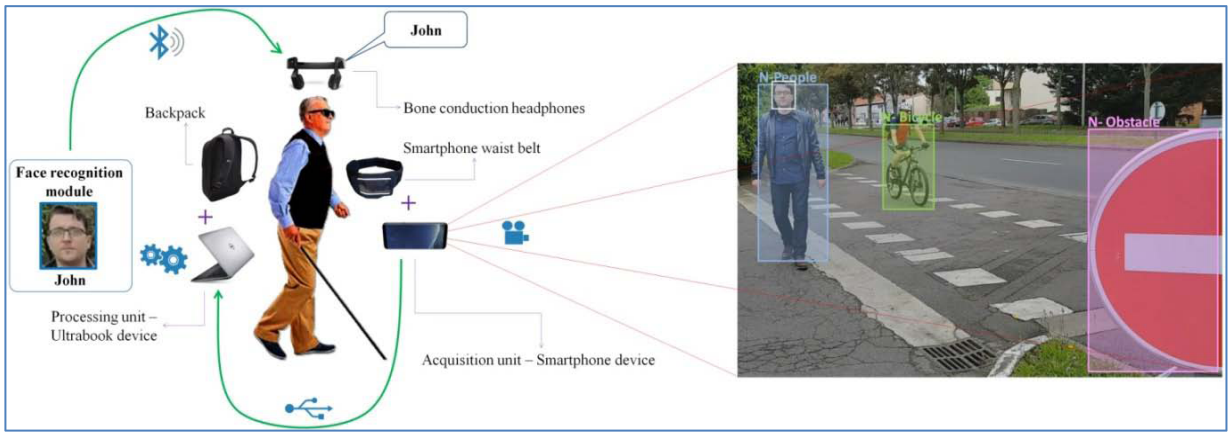

FIGURE 1. The hardware architecture of the proposed ***DEEP-SEE FACE*** system.

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.

The framework has been extended in [23], where authors propose a CNN-based approach to perform both people detection and recognition. Even though the method returns good results for the detection module the performance of the recognition system is inferior to 70% and is influenced by lighting condition or by user/camera motion. In addition, the system has never been tested with actual visually impaired people and nothing is said about the hardware architecture or about the acoustic warning messages.

The SmartCane face recognition system dedicated to blind people is introduced in [12]. The framework functions in real-time and is designed to identify persons around the VI, while informing the user about their presence through a set of vibration patterns. The face detection algorithm is based on Adaboost, while for recognition the compressed sensing with L2 norm classifier is used. However, because the video camera needs to be head-worn the framework is considered invasive.

In [24], a prototype that helps the VI people to interact with other humans is introduced. The system uses a regular smartphone device in conjunction with a wireless network in order to detect and recognize people standing in front of the VI user. The warning messages are transmitted through a set of acoustic patterns. However, despite the efficient recognition scores reported (superior to 96%), the system was tested solely in simulated, indoor scenarios with less than ten people in the recognition database.

The Facial Expression Perception through Sound (FEPS) sensorial substitution system is proposed in [25]. The system is designed to improve the VI people participation in social communication by perceiving the interlocutor's facial expression. Even though the project's goals are ambitious, the accuracy of the system is relatively low and the computational time is extensive.

Recently, in [26], a real-time face recognition system that combines face matching and identity verification is proposed. By exploiting the temporal efficiency of matching and a traditional classifier (SVM), the system is able to inform a VI user about the presence of a known identity in the near surroundings. Even though the system is designed to work in real-time on a computer with relatively reduced processing capabilities, the framework has never been tested with real VI users or in outdoor scenarios.

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*** 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.

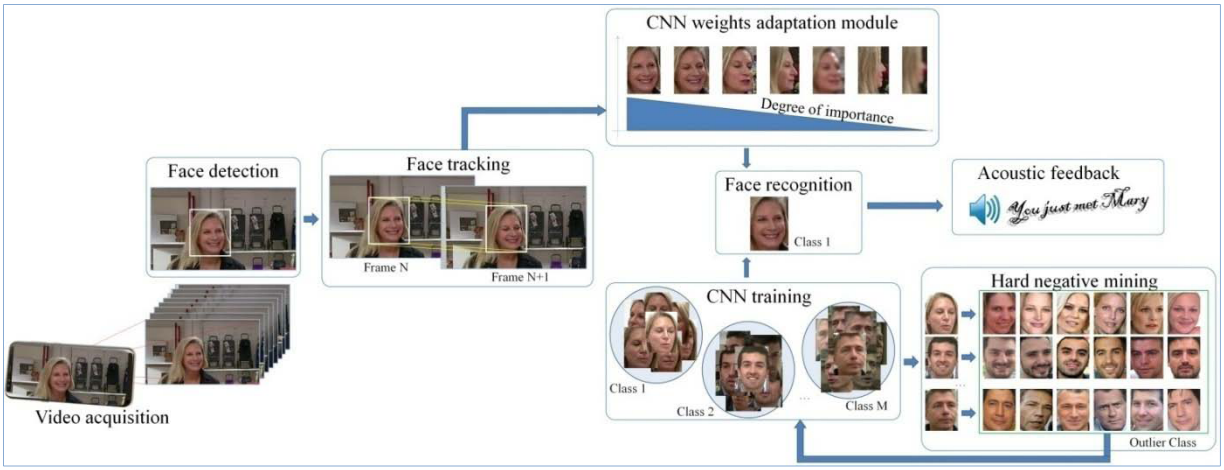

FIGURE 2. The proposed ***DEEP-SEE FACE*** methodological framework.

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 [28] 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.

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).

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.

The VGG output is a 4096-dimensional feature vector representation (corresponding to the penultimate layer) of the face, which is further normalized to a unit vector.

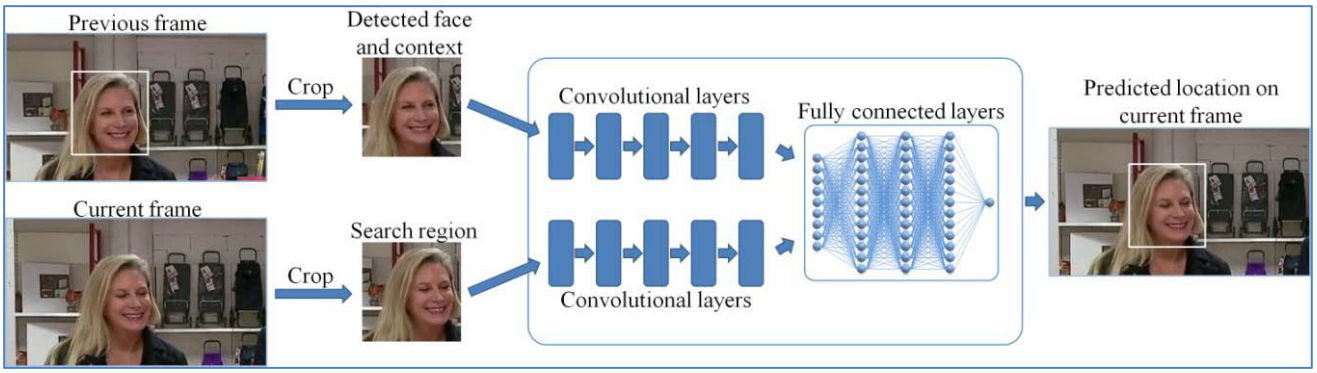

FIGURE 3. Face tracking using a modified version of the ATLAS algorithm adapted to the scenario of multiple face tracking.

Each feature face representation is further fed to a weight adaptation scheme, described in the following paragraphs.

Given a face that is tracked in successive frames of a video stream, the face recognition module is designed to determine the probability of a face to belong to a specific category.

Let us denote by $F = \{x_1, x_2, \ldots, x_L\}$ a face tracked in a video sequence of length L frames, where $x_k, k = 1, \ldots, L$ is a face instance in the k^{th} frame of the considered video. At each frame, the considered face x_k has its corresponding normalized feature representation f_k that is extracted from the VGG16 module.

Our objective is to create a global descriptor, denoted by $r(F)$ and associated to face F that aggregates all the features extracted from multiple video frames (and which correspond to multiple face instances) into a compact, global face representation, defined as:

$$r(F) = \sum_k w_k \cdot f_k, \tag{1}$$

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.

The key ingredient in equation (1) is the set of weights w_k. A simple approach, as the one introduced in [33], would consist in a naive averaging, which corresponds to equal weights $w_k = 1/L$. However, such an approach is not optimal, because all face instances are treated with equivalent importance.

In this paper, we have designed a learning-based optimized scheme, described in the following section that adaptively modifies the scores depending on the degree of noise within the frame, face poses or viewing angles.

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 overfitting 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.

For the linear motion blur, as in [38], we used a kernel length that is randomly selected within the [5], [15] interval and a kernel angle ranging between 10 and 30 degrees. For scale variation, we have considered various down-sampling factors between 1/12 and 1/2 of the original image size. Finally, the face instances have been compressed using the JPEG compression algorithm at a quality parameter randomly selected within the [10], [50] interval. At the end, we have obtained a database of about 1Mimages.

The weight adaptation module receives all features and generates the corresponding weights for them. Specifically, for the f_k face feature vector, the output is a value that corresponds to the face significance s_k, which represents the probability to belong to the *relevant* category issued by the VGG network. Finally, the s_k coefficients are passed through a softmax operator to obtain the weights w_k with $\sum_k w_k = 1$:

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

In this way, we ensure the robustness of our approach that is invariant to the number of face instances (that can vary from person to person) or to the order it receives the images (the global face descriptor will be the same regardless if the face instances are reversed or reshuffled).

Figure 4 presents some examples of weights computed using our adaptation module on various videos: (1) 5 video recorded in real urban scenes by actual VI users and (2) 25 image sequence selected from the France national television broadcast. As it can be observed, blurred, partially occluded or profile face instances play a reduced role in the global, aggregated face descriptor that is further used for classification purposes.

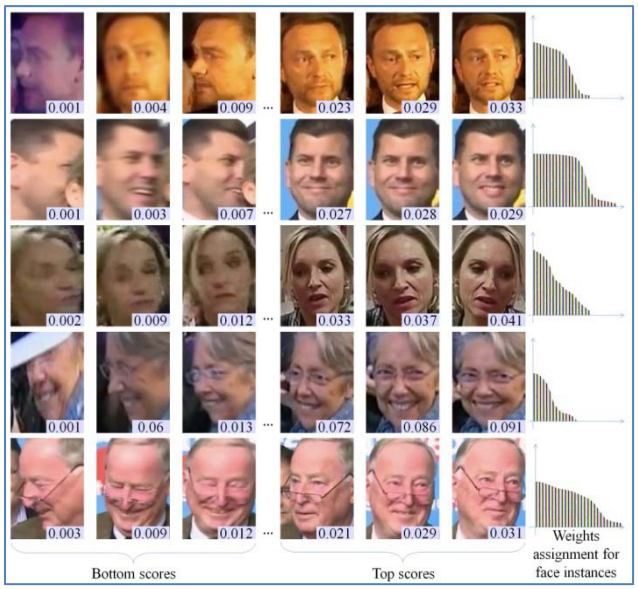

FIGURE 4. Visual examples of face instances and their associated weights displayed in ascending order with respect to the video content variation.

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 ensure a better degree of generalization for the *"Outlier"* class. The faces included in the negative examples category have been selected from the extended version of the Multi-Task Facial Landmark (MTFL) dataset as presented in Section III.C.1. In the experimental evaluation section, we also analyze the impact of the parameter N over the system's performance.

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.

IV. 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 1280×720 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 [33] 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 $\sim$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

TABLE 1. Experimental results of the ***DEEP-SEE FACE*** recognition module.

Method	Ground Truth (Tracked faces / Known identities)	True Positive (TP)	False Positive (FP)	False Negative (FN)	Accuracy (%)	Recognition Rate (%)	F1 score (%)
(1). Frame-based method	6214 / 1108	800	401	308	66.61	72.21	69.29
(2). Baseline aggregation method		912	264	196	77.55	82.31	79.85
(3)Weight adaptation method		983	215	125	82.05	88.71	85.25
(4). Weight adaptation with random "Outliers"		1017	142	91	87.74	91.78	89.72
(5) Weight adaptation with hard negative mining for the "Outlier" category		**1044**	**86**	**64**	**92.38**	**94.22**	**93.29**

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.

The evaluation of the proposed face recognition system is performed using traditional objective parameters such as Accuracy (A), Recognition rate (R) and F1 norm, defined as described in equation (3):

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

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).

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.

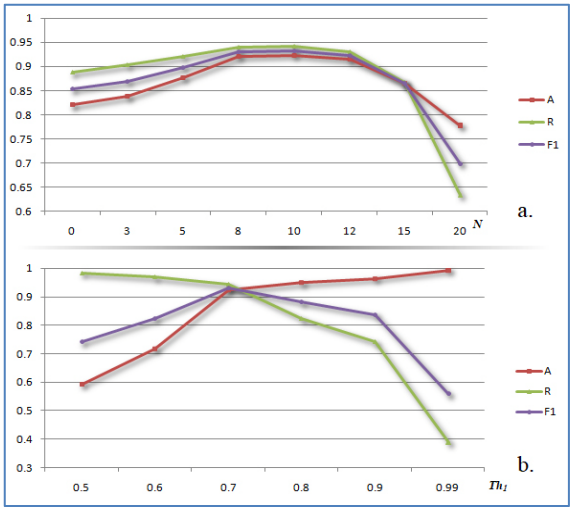

FIGURE 5. The system performance variation with the different parameters involved. (a) The first N hardest negative examples; (b) The probability threshold (Th_1) of assigning a face to a specific class.

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

(1) A per-frame approach that applies the face recognition algorithm to each individual frame and then takes a decision based on the dominant class;

(2) 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];

(3) A compact face representation method that for each face tracked between successive frames uses a weight adaptation method as presented in Section III.C;

(4) A face recognition module that contains both the weight adaptation scheme and an *"Outlier"* class constructed with randomly selected samples.

(5) 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 1.

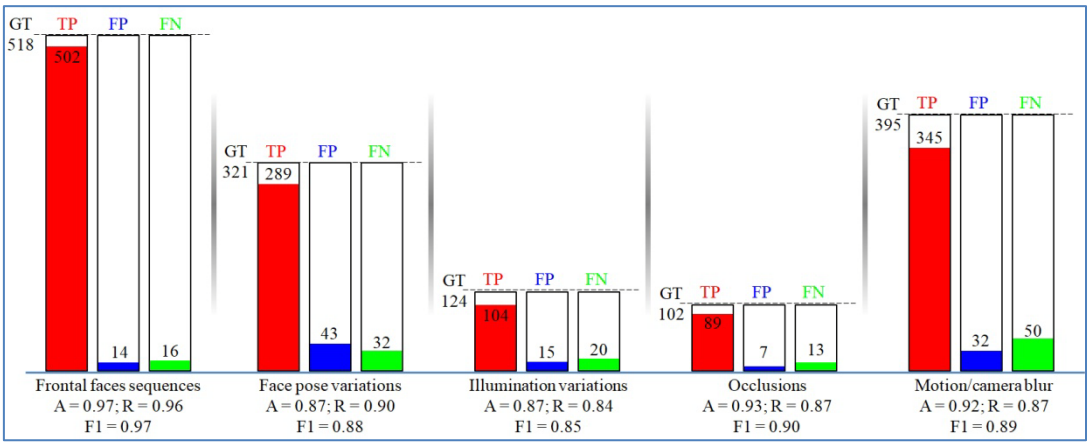

FIGURE 6. ***DEEP-SEE FACE*** **performance evaluation with various indoor/outdoor conditions.**

From the results presented in Table 1 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. Treating each frame as an independent image makes it impossible to differentiate between discriminative and poor face instances. (2) The average-based aggregation of the face features representation improves the recognition scores with more than 10% gain in F-score. However, it is obvious that high-quality frames return higher recognition scores than low-quality ones. The results obtained by the adaptive weighting scheme clearly show that such an approach is more appropriate with an F1-score of 85.25%. Finally, we can observe that when unknown face identities are applied as input it is important to develop an "*Outlier*" category that significantly reduces the number of false alarms. In this case, the F1-score are significantly increasing, with 89.72% for the random selection strategy and 93% for the hard mining approach.

From the computational point of view, in order to ensure real-time performances, the system performs an initial assignment to a category as soon as more than 10 face frames are included in the ***relevant*** class, and not when the face tracking is completed.

Concerning the 30 videos acquired by VI users, they led to a total number of 1108 known face instances (frames). In the results presented in Table 1, we have considered for classification purposes only faces situated at a distance inferior to 5 meters relative to the video camera attached to the VI user. In this context, we have constrained the face size applied as input to the recognition module to have a resolution superior to 64×64 pixels.

The 1108 known faces 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. In Figure 6, we present the obtained performances on each of the considered category.

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. The lowest performances are obtained for blurred face instances, while the highest scores are obtained for frontal faces.

From a practical point of view, the battery usage of the proposed hardware architecture is one of the most important parameters that need to be taken into account. First, the video camera embedded on the smartphone is used as an acquisition device that constantly records the surrounding scene and transmits the video stream to the processing unit situated on the VI user backpack. Second, the ultrabook computer processes the acquired frames and applies the face recognition method that is computationally expensive. Third, the continuous connection between the smartphone and the processing unit drains the system energy. Finally, the feedback provided to the VI user through bone conduction headphones also consumes energy.

After analyzing the lifetime of our application when integrated in the ***DEEP-SEE*** framework we observed that the system can function continuously for more than 2 hours without the need of an additional recharging of any of the components. However, a major drawback of using a backpack computer as processing unit is the ultrabook heating when performing all computations on the GPU boards.

In terms of computational speed, when implementing the whole framework on a regular ultrabook computer having Linux Ubuntu (version 16.04) as operating system, with 32 GB of RAM, i7-7700 CPU at 2.8 GHz and running on an NvidiaGTX 1050 GPU (768 cores and 4GB frame buffer), CUDA version 9.2the average processing speed is around 4 - 5 fps.

D. SUBJECTIVE SYSTEM EVALUATION

The qualitative system evaluation was performed with the help of a group of 5 actual visually impaired people with ages

ranging between 25 and 65 years. The goal of the evaluation was to determine if: (1) the users were able to start the ***DEEP-SEE FACE*** framework by their own, (2) the users are informed about the presence of a novel person within the scene using the proposed acoustic signals and (3) the global framework is useful to complement the white cane. The tests have been performed in various indoor and outdoor environments for which the VI people had no initial knowledge about which familiar persons were present. After our discussions with the participants, the following conclusions can be highlighted:

(1). The VI users have found the system friendly and wearable, satisfying the hands-free and ears-free conditions imposed by the blind community on any assistive device.

(2). The VI people, familiar with handling smartphones, manifested a strong interest in the architecture even in the presence of some classification errors.

(3) Some partially sighted people expressed interest in using our system in daily activities and consider the proposed acoustic warnings intuitive.

(4) At the beginning of the testing phase, we have observed the retention and mistrust to innovations, especially for older people. However, after short training stage, all participants declared that the framework is easy to learn and useful.

(5) The proposed system should be used to complement the white cane with additional functionalities, because most of the VI people do not feel confident enough to use the ***DEEP-SEE*** framework as a standalone assistive device.

V. CONCLUSION AND PERSPECTIVES

In this paper we have introduced a face-recognition assistive device so-called ***DEEP-SEE FACE***, designed to improve cognition of visually impaired people when interacting with other persons in social encounters.

The proposed approach does not require any *a priori* knowledge about the position of various people existent in the scene and jointly exploits computer vision algorithms and deep convolutional neural networks (CNNs) in order to improve cognition of VI users. By using the VGG CNNs architecture combined with region proposal framework the system that receives as input the entire video frame is able to correctly detect, track and recognize, in real-time various persons situated at arbitrary locations.

The semantic interpretation of the recognized person identity is transmitted to the VI user as a set of acoustic warnings.

From the methodological point of view, the core of the approach relies on a novel video-based face recognition framework able to construct an effective global, fixed-size face representation method, 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. Secondly, a hard negative mining stage is proposed that helps us differentiate between known and unknown face identities.

The experimental evaluation performed on a large dataset of 30 videos acquired with the help of VI people validate the proposed methodology, which is able to return a recognition rate superior to 92% regardless on the lighting conditions, face pose or various types of motion existent in the scene.

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.

Moreover, when looking at the emerging trends in the smartphone industry, we can observe that various constructors begin to propose hardware prototypes dedicated to CNN applications. Within this context, let us mention the artificial intelligence chips recently launched by CEVA (*e.g.*, NP4000) or Samsung (*e.g.*, Exynos 9 Series 9810) at the Consumer Electronics Symposium (CES'2018). We hope that such technologies will permit us, in the recent future, to autonomously run the ***DEEP SEE FACE*** framework on a smartphone device.

REFERENCES

[1] J. L. Obermayer, W. T. Riley, O. Asif, and J. Jean-Mary, "College smoking-cessation using cell phone text messaging," *J. Amer. College Health*, vol. 53, no. 2, pp. 71–78, 2004.

[2] S. Haug, C. Meyer, G. Schorr, S. Bauer, and U. John, "Continuous individual support of smoking cessation using text messaging: A pilot experimental study," *Nicotine Tobacco Res.*, vol. 11, no. 8, pp. 915–923, 2009.

[3] D. Scherr, R. Zweiker, A. Kollmann, P. Kastner, G. Schreier, and F. M. Fruhwald, "Mobile phone-based surveillance of cardiac patients at home," *J. Telemedicine Telecare*, vol. 12, no. 5, pp. 255–261, 2006.

[4] P. Rubel *et al.*, "Toward personal eHealth in cardiology. Results from the EPI-MEDICS telemedicine project," *J. Electrocardiol.*, vol. 38, no. 4, pp. 100–106, 2005.

[5] S. C. Wangberg, E. Årsand, and N. Andersson, "Diabetes education via mobile text messaging," *J. Telemed. Telecare*, vol. 12, no. 1, pp. 55–56, 2006.

[6] P. Mohan, D. Marin, S. Sultan, and A. Deen, "MediNet: Personalizing the self-care process for patients with diabetes and cardiovascular disease using mobile telephony," in *Proc. IEEE 30th Annu. Int. Conf. Eng. Med. Biol. Soc.*, Aug. 2008, pp. 755–758.

[7] M. Jones, J. Morris, and F. Deruyter, "Mobile healthcare and people with disabilities: Current state and future needs," *Int. J. Environ. Res. Public Health*, vol. 15, no. 3, p. 515, 2018.

[8] *A World Health Organization (WHO)—Visual Impairment and Blindness*. Accessed: Jul. 5, 2018. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs282/en/

[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*, vol. 12, no. 12, pp. 17476–17496, 2012, doi: 10.3390/s121217476.

[10] R. Tapu, B. Mocanu, and T. Zaharia, "*DEEP-SEE*: Joint object detection, tracking and recognition with application to visually impaired navigational assistance," *Sensor*, vol. 17, no. 11, p. 2473, 2017, doi: 10.3390/s17112473.

[11] S. Chaudhry and R. Chandra, "Design of a mobile face recognition system for visually impaired persons," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1502.00756

[12] Y. Jin, J. Kim, B. Kim, R. Mallipeddi, and M. Lee, "Smart cane: Face recognition system for blind," in *Proc. 3rd Int. Conf. Hum.-Agent Interact. (HAI)*, New York, NY, USA: 2015, pp. 145–148 .

[13] Y. Rao, J. Lu, and J. Zhou, "Attention-aware deep reinforcement learning for video face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3951–3960.

[14] R. Tapu, B. Mocanu, A. Bursuc, and T. Zaharia, "A smartphone-based obstacle detection and classification system for assisting visually impaired people," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Sydney, NSW, Australia, Dec. 2013, pp. 444–451.
[15] R. Manduchi, "Mobile vision as Assistive technology for the blind: An experimental study," in *Computers Helping People With Special Needs*, 2012, pp. 9–16.
[16] J. M. H. D. Buf *et al.*, "The smartvision navigation prototype for the blind," in *Proc. Int. Conf. Softw. Develop. Enhancing Accessibility Fighting Info Exclusion (DSAI)*, 2010, pp. 167–174.
[17] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna, "Face recognition based on Kinect," *Pattern Anal. Appl.*, vol. 19, no. 4, pp. 977–987, 2016.
[18] J. B. C. Neto and A. N. 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. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2013, pp. 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.*, Sep./Oct. 2013, pp. 1–6.
[21] S. Berretti, N. Werghi, A. del Bimbo, and P. Pala, "Selecting stable keypoints 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," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 1, pp. 52–64, Feb. 2017.
[23] S. Chaudhry and R. Chandra, "Face detection and recognition in an unconstrained environment for mobile visual aSSISTIVE system," *Appl. Soft Comput.*, vol. 53, pp. 168–180, Apr. 2017, doi: 10.1016/j.asoc.2016.12.035.
[24] K. M. Kramer, D. S. Hedin, and D. J. Rolkosky, "Smartphone based face recognition tool for the blind," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol.*, Aug./Sep. 2010, pp. 4538–4541.
[25] M. I. Tanveer, A. S. M. I. Anam, A. K. M. M. Rahman, S. Ghosh, and M. Yeasin, "FEPS: A sensory substitution system for the blind to perceive facial expressions," in *Proc. 14th Int. ACM SIGACCESS Conf. Comput. Accessibility*, New York, NY, USA, 2012, pp. 207–208.
[26] G. Fusco, N. Noceti, and F. Odone, "Combining retrieval and classification for real-time face recognition," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Beijing, China, Sep. 2012, pp. 276–281.
[27] G. Gou, Z. Li, G. Xiong, Y. Guan, and J. Shi, "Video face recognition through multi-scale and optimization of margin distributions," in *Proc. Int. Conf. Comput. Sci. (ICCS)*, Zürich, Switzerland, vol. 108, 2017, pp. 2458–2462.
[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
[29] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May/Jun. 2017, pp. 650–657, doi: 10.1109/FG.2017.82.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556
[31] B. Mocanu, R. Tapu, and T. Zaharia, "Single object tracking using offline trained deep regression networks," in *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov./Dec. 2017, pp. 1–6, doi: 10.1109/IPTA.2017.8310091.
[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 448–456.
[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, vol. 1, no. 3, p. 6.
[34] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. ECCV*, 2014, pp. 94–108.
[35] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image Vis. Comput.*, vol. 47, pp. 27–35, Mar. 2016.
[36] 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, Apr. 2009.
[37] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2012, pp. 1097–1105.
[38] K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, and M. Chandraker, "Unsupervised domain adaptation for face recognition in unlabeled videos," in *Proc. ICCV*, 2017, pp. 3210–3218.
[39] D. Mishkin, N. Sergievskiy, and J. Matas. (2016). "Systematic evaluation of CNN advances on the ImageNet." [Online]. Available: https://arxiv.org/abs/1606.02228
[40] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell, "Best practices for fine-tuning visual classifiers to new domains," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 435–442. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-49409-8_34
[41] J. Yang *et al.*, "Neural aggregation network for video face recognition," in *Proc. 32nd IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5216–5225.

BOGDAN MOCANU received the B.S. degree in electronics, telecommunications, and information technology and the Ph.D. degree in electronics and telecommunication from the University Politehnica of Bucharest, Romania, in 2008 and 2011, respectively, and the second Ph.D. degree in informatics from University Paris VI–Pierre et Marie Currie, Paris, France, in 2012. Since 2012, he has been a Researcher with the ARTEMIS Department, Institut Mines-Télécom/Télécom SudParis, France. His major research interest is computer application technology: such as 3-D model compression and algorithm analysis in image processing.

RUXANDRA TAPU received the B.S. degree (Valedictorian) in electronics, telecommunications, and information technology and the Ph.D. degree in electronics and telecommunication from the University Politehnica of Bucharest, Romania, in 2008 and 2011, respectively, and the second Ph.D. degree (Hons.) in informatics from University Paris VI–Pierre et Marie Currie, Paris, France, in 2012. Since 2012, she has been a Senior Researcher with the ARTEMIS Department, Institut Mines-Télécom/Télécom SudParis, France, having as major research interest content-based video indexing and retrieval, pattern recognition, and machine learning techniques.

TITUS ZAHARIA (M'97) received the Engineer degree in electronics and telecommunications and the M.Sc. degree from the Politehnica University of Bucharest, Bucharest, Romania, in 1995 and 1996, respectively, and the Ph.D. degree in mathematics and computer science from University Paris V—Rene Descartes, Paris, France, in 2001. He joined the ARTEMIS Department, Institut Télécom, Télécom SudParis, as an Associate Professor, in 2002, and has become a Full Professor in 2011. His research interests include visual content representation methods, with 2-D/3-D compression, reconstruction, recognition, and indexing applications. Since 1998, he actively contributes to the ISO/MPEG-4 and MPEG-7 standards.