CHAPTER 1

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 monitor patients with chronic heart failure detect early signs of arrhythmia or ischemia, provide diabetes education or monitor relevant physiological markers . can effectively link disparate Bitcoin transactions to a common user and, in many cases, to that user's real-world identity.

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 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. Here 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) 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. 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.

So, 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 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.

CHAPTER 2

SYSTEM ARCHITECTURE

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, Mobile Vision and Smart Vision 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.*, Cardia Neto and Maran, Li *et al.*, Goswami *et al.* and Berretti *et al.* 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. 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 and Berretti *et al.* .

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. 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, 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 Smart Cane face recognition system dedicated to blind

people is introduced in. 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. This is 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.

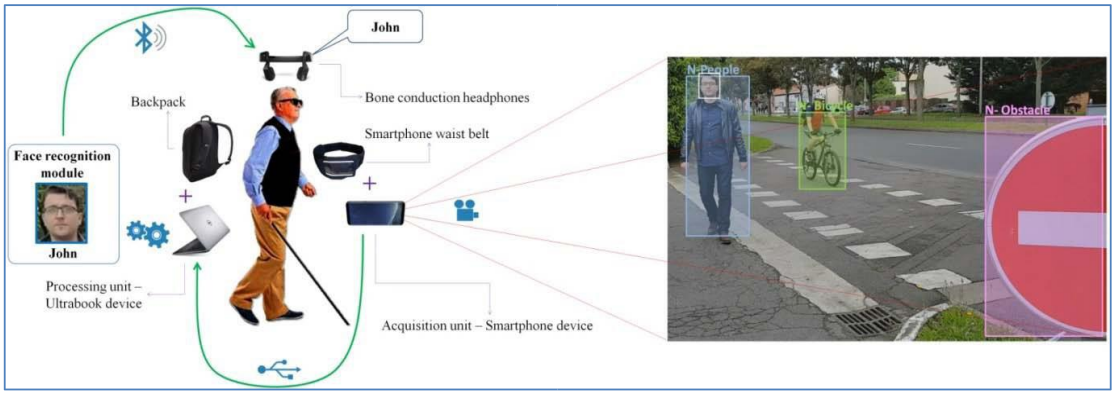


Figure 1. The hardware architecture of the proposed *DEEP-SEE FACE* system.

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.

The Facial Expression Perception through Sound (FEPS) sensorial substitution system is proposed(Figure 1). 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, 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, 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.

CHAPTER 3

TECHNOLOGY USED

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 × H (where W and H represent the width and height, respectively), we obtain a maximum number of W ×H×n proposals. As indicated, 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. 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.

FACE TRACKING

The tracking system takes as input, at a given frame, the face bounding box indicated by the detection module. Then, the goal is to determine the face position between consecutive frames. The tracking methodology is based on our previous ATLAS algorithm introduceduses accounts. Like bank account funds, ether tokens appear in a wallet, and can be ported (so to speak) to another account. Proposed system 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. 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.

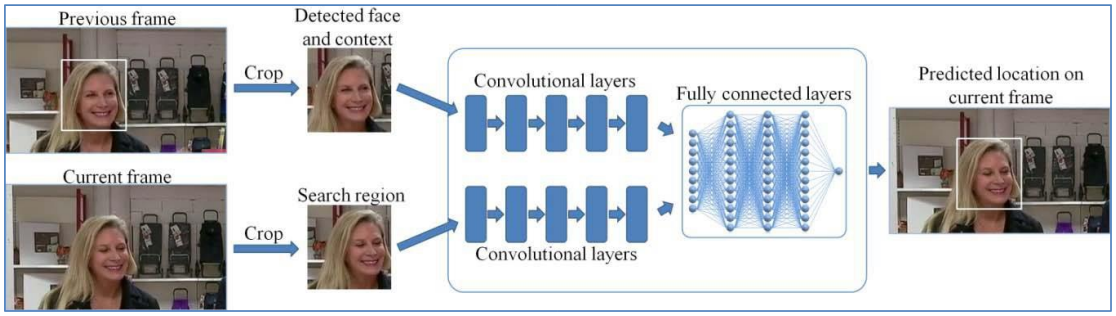


Figure 2: Face tracking using a modified version.

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 network architecture with the batch normalization strategy introduced (Fig 2). 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.

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:

$$r(F) = \Sigma_k \, w_k \cdot f_k \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, (1)$$

where $\{w_k\}^{L}_{k=1}$ 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 singleframe face representation. The key ingredient in equation (1) is the set of weights w_k . A simple approach, as the one introduced earlier, 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.

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 , 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.

In order to determine the blurriness degree of the considered faces, we have adopted a non-referential sharpness (NRS) metric that determines the local contrast in the neighbourhood 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.

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 $\Sigma_k \, w_k = 1$:

$$w_k = \exp(s_k) / \Sigma_j \exp(s_j) \quad , (2)$$

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).

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 returns false alarms. These non-face regions should be also marked as unknown instances.

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 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.

CHAPTER 4

IMLEMENTATION

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.

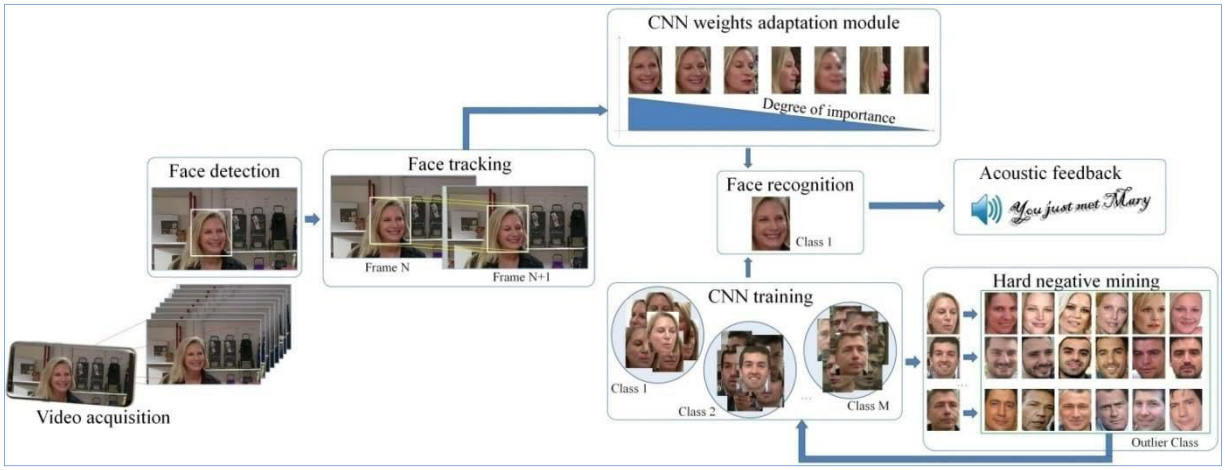


Figure 3: The proposed DEEP-SEE FACE methodological framework.

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 and aligned using the facial landmarks. 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.

QUANTITATIVE SYSTEM EVALUATION

The proposed face recognition system was tested on the set of 30 video streams . Because the image sequences were recorded either in crowded urban scenes or in studio with audience, more than 5.000 unknown individual were identified in the videos. In addition, the same person may appear in various environments, while in the same location various people may be present. In the evaluation, the testing dataset is different from the face instances used for training. 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):

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

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).

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**

Table 1: Experimental results of the DEEP-SEE FACE recognition module

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.

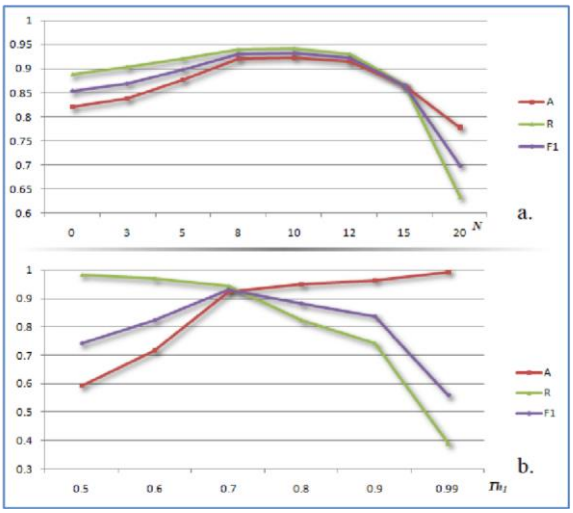


Figure 4: . The system performance variation with the different parameters involved. (a) The first N hardest negative examples; (b) The probability threshold (Th1) of assigning a face to a specific class.

Figure 4 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 Th1 parameter is fixed to 0.7.

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;

(3) A compact face representation method that for each face tracked between successive frames uses a weight adaptation method;

(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

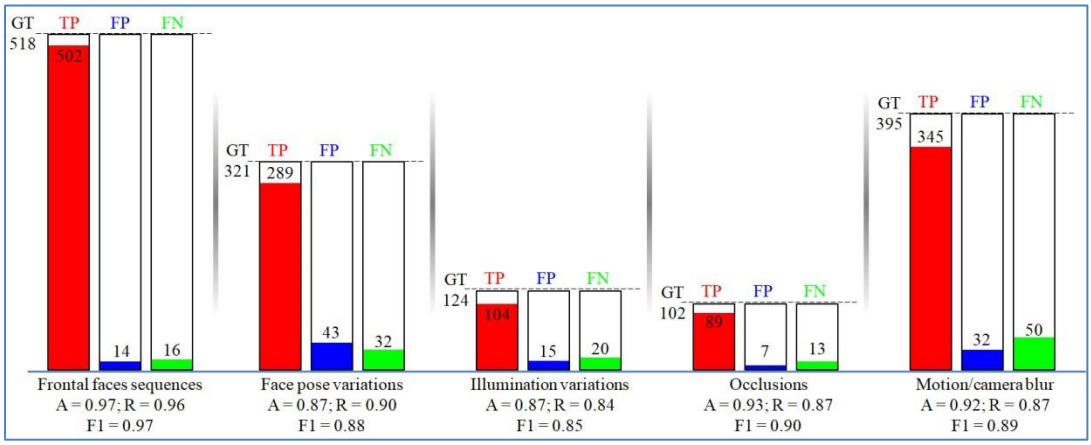


Figure 5. DEEP-SEE FACE performance evaluation with various indoor/outdoor conditions

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.

CHAPTER 5

APPLICATIONS

Health risk assessment and biomarkers of aging discovery

CNNs can be naturally tailored to analyze a sufficiently large collection of time series data representing one-week-long human physical activity streams augmented by the rich clinical data (including the death register, as provided by, e.g., the NHANES study). A simple CNN was combined with Cox-Gompertz proportional hazards model and used to produce a proof-of-concept example of digital biomarkers of aging in the form of all-causes-mortality predictor.

Image recognition

CNNs are often used in image recognition systems. In 2012 an error rate of 0.23 percent on the MNIST database was reported.Another paper on using CNN for image classification reported that the learning process was "surprisingly fast"; in the same paper, the best published results as of 2011 were achieved in the MNIST database and the NORB database. Subsequently, a similar CNN called AlexNet won the ImageNet Large Scale Visual Recognition Challenge 2012.

When applied to facial recognition, CNNs achieved a large decrease in error rate. Another paper reported a 97.6 percent recognition rate on "5,600 still images of more than 10 subjects".CNNs were used to assess video quality in an objective way after manual training; the resulting system had a very low root mean square error.

The ImageNet Large Scale Visual Recognition Challenge is a benchmark in object classification and detection, with millions of images and hundreds of object classes. In the ILSVRC 2014,a large-scale visual recognition challenge, almost every highly ranked team used CNN as their basic framework. The winner GoogLeNet(the foundation of DeepDream) increased the mean average precision of object detection to 0.439329, and reduced classification error to 0.06656, the best result to date. Its network applied more than 30 layers. That performance of convolutional neural networks on the ImageNet tests was close to that of humans. The best algorithms still struggle with objects that are small or thin, such as a small

ant on a stem of a flower or a person holding a quill in their hand. They also have trouble with images that have been distorted with filters, an increasingly common phenomenon with modern digital cameras. By contrast, those kinds of images rarely trouble humans. Humans, however, tend to have trouble with other issues. For example, they are not good at classifying objects into fine-grained categories such as the particular breed of dog or species of bird, whereas convolutional neural networks handle this.

In 2015 a many-layered CNN demonstrated the ability to spot faces from a wide range of angles, including upside down, even when partially occluded, with competitive performance. The network was trained on a database of 200,000 images that included faces at various angles and orientations and a further 20 million images without faces. They used batches of 128 images over 50,000 iterations.

Natural language processing

CNNs have also been explored for natural language processing. CNN models are effective for various NLP problems and achieved excellent results in semantic parsing, search query retrieval, sentence modeling, classification, prediction and other traditional NLP tasks.

CHAPTER 6

FUTURE SCOPE

For further work and developments and 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.

Another way that innovators are looking to implement facial recognition is within subways and other transportation outlets. They are looking to leverage this technology to use faces as credit cards to pay for your transportation fee. Instead of having to go to a booth to buy a ticket for a fare, the face recognition would take your face, run it through a system, and charge the account that you've previously created. This could potentially streamline the process and optimize the flow of traffic drastically.

Face recognition systems used today work very well under constrained conditions, although all systems work much better with frontal mug-shot images and constant lighting. All current face recognition algorithms fail under the vastly varying conditions under which humans need to and are able to identify other people. Next generation person recognition systems will need to recognize people in real-time and in much less constrained situations

CONCLUSION

With the introduction of 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.

BIBLIOGRAPHY

[1] J. L. Obermayer, W. T. Riley, O. Asif, and J. Jean-Mary, ‘‘College smokingcessation 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 .

APPENDIX

Global Academy of Technology

Department of Computer Science and Engineering

Technical Seminar (15CSS86)

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

Name: Sathya Narayana R
USN: 1GA15CS134

Guide: Krishna Prasad R
Assistant Professor
Dept. of CSE, GAT

Index

Dept. of CSE
15CSS86: Technical Seminar
2

Objectives

- **Neural network -** deep convolutional neural networks in order to detect, track, and recognize, in real-time, various persons existent in the video streams.
- **Weight adoption scheme -** weight adaptation scheme 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.
- **Negative mining -** negative mining stage that helps us differentiating between known and unknown face identities.
- **Alerts -** the semantic information about the presence of a familiar is delivered with the help of acoustic warning messages, transmitted through bone conduction headphones.

Dept. of CSE
15CSS86: Technical Seminar
3

Abstract

In recent years, several frameworks based on mobile platforms and dedicated to healthcare services have emerged. Nowadays, more than 285 million people worldwide suffer from visual impairment.

The visually impaired people adapt to normal life by using traditional assistive aids, such as white canes or walking dogs.

The white cane is widely accepted by the blind community. However, such an assistive element shows quickly its limitations when confronted with the high diversity of situations

Dept. of CSE
15CSS86: Technical Seminar
4

Introduction

- DEEP-SEE FACE, a assistive device based on computer vision algorithms and offline-trained deep convolutional neural network that extends the 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.

Dept. of CSE
15CSS86: Technical Seminar
5

Technologies Used

- CNN
- Let us denote by F D fx1; $x2$; : : : ; xLg a face tracked in a video sequence of length L frames, where xk ; k D 1; : : : ; L is a face instance in the kth frame of the considered video.
- global face representation,defined as:

$$r(F) = \Sigma_k w_k . f_k$$

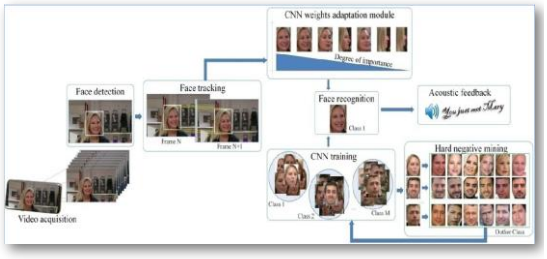
- where $\{w_k\}^L_{k=1}$ is a set of weights, with w_k the coefcient associated to the feature of the kth frame.

Dept. of CSE
15CSS86: Technical Seminar
6

Implementation

- Face detection
- Face tracking
- Face recognition
- Weight adoption module
- Hard negative mining
- Acoustic feedback

Implementation



Issues

- 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.
- So, two challenges are addressed:

1 detect and recognize familiar people when navigating in indoor or outdoor environment

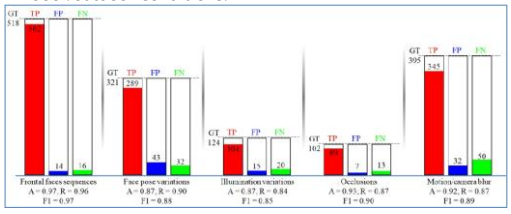
2 acquire additional information about the identity of various people/celebrities appearing on the media broadcasted at TV or over internet.

Application

- The state-of-the-art approaches dedicated to the VI assistive devices based on computer vision/machine learning methods.
- It is possible to obtain high recognition rates on mobile wearable devices.
- This system does not require any dedicated hardware architecture and can be accessible to any VI user at low cost.

Comparison Results

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



Conclusion & Future Scope

- 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.
- 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.
- we can observe that various constructors begin to propose hardware prototypes dedicated to CNN applications.

Bibliography

[1] 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. 255261, 2006.

[2] 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.

[3] A World Health Organization (WHO)Visual Impairment and Blind-ness. Accessed: Jul. 5, 2018.Available:http://www.who.int/mediacentre/factsheets

[4] 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.

[5] 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.

Thank You