Dependable Person Recognition by Means of Local Descriptors of Dynamic Facial Features

Aniello Castiglione[1], Giampiero Grazioli[2], Simone Iengo[2], Michele Nappi[2], and Stefano Ricciardi[3(✉)]

[1] Department of Science and Technology, University of Naples Parthenope, Naples, Italy
castiglione@ieee.org, castiglione@acm.org
[2] Department of Computer Science, University of Salerno, Fisciano, Italy
{g.grazioli,s.iengo1}@studenti.unisa.it, mnappi@unisa.it
[3] Department of Biosciences, University of Molise, Campobasso, Italy
stefano.ricciardi@unimol.it

Abstract. In this work, a complementary approach that adds a dynamic component to face biometrics is proposed. The dynamic appearance and the time-dependent local features characterizing the face of an individual during speech utterance are indeed considered in their spatial and temporal components. Ultimately, the aim is to capture, represent and compare facial patterns related to speech utterance, to improve biometric system dependability thanks to an intrinsically difficult to forge descriptor. The proposed approach applies the concept of dynamic texture to the domain of person identification through dynamic facial patterns modeled by means of the Volume Local Binary Pattern (VLBP) descriptor, which effectively combines local features and movement. To the aim of improving the efficiency of this technique, only the occurrences of the Local Binary Patterns related to Three Orthogonal Planes (LBP-TOP) have been considered. A deep feed forward network has been trained and optimized on video samples from the XM2VTS database concerning utterance of a given sentence. The results obtained in the recognition task performed on test video sequences confirm that the proposed approach features state-of-the-art performances with regard to accuracy and robustness of the identification.

Keywords: Biometrics · Face recognition · Image analysis · Face biometrics · Dependability

1 Introduction

Dependability of biometric systems is a key aspect in their worldwide diffusion and everyday usage, regardless of the specific application they are supposed to improve, and it is tightly related to the overall reliability in the process of

G. Wang et al. (Eds.): DependSys 2019, CCIS 1123, pp. 247–261, 2019.
https://doi.org/10.1007/978-981-15-1304-6_20

accessing a given resource or a given place. Nevertheless, it is worth noting that while a given biometric system could perform well in terms of accuracy and robustness (i.e. featuring low False Acceptance Rate and high False Rejection Rate), this does not automatically means it is dependable [24]. Dependability of a biometric system, indeed, implies much more than a high performance of the processing pipelines (though the latter is a fundamental requirement), since it involves other aspects such as the reliability of the capture process, the capability to cope with uncontrolled conditions and, not secondarily, the resistance to attacks from malicious users.

In this work we focus on this last aspect of the dependability of a biometric system, with particular regard to face biometrics which represents one of the most diffused way to perform person authentication and identification in a contactless and natural way. The idea inspiring the proposed approach is to increase the level of resistance of face biometrics to presentation attacks [4], by exploiting face dynamics related to utterance of a given sentence. These dynamic facial features are subject dependent and represent a sort of motion signature involving much greater difficulty in counterfeiting it, compared to static face representations. To this aim, video capture of face changes during sentence utterance (see Fig. 1) are used to extract dynamic local features from face lower half, by means of the LBP-TOP variant of the Volume Local Binary Pattern (VLBP) method. These spatial-temporal features are then used to train a deep feed-forward neural network and subsequently to find correspondences between the probe descriptor and the available gallery. The experiments conducted on audiovisual samples of the public XM2VTS dataset show state-of-the-art recognition accuracy exceeding 99%, along with a high robustness to intra-class variations (the way sentence is pronounced by the same subject) and good independence from the choice of the sentence, confirming the advantages of using inherently dependable dynamic facial features.

The rest of this paper is organized as follows: Sect. 2 resumes a selection of works related to the present study; Sect. 3 presents in detail the proposed approach to inherently safer face biometrics; Sect. 4 describes the results from the experiments carried out. Finally, Sect. 5 draws conclusions, along with directions for future research.

2 Related Works

Our proposal is aimed at extracting, representing and matching dynamic facial features related to the way a sentence is pronounced. Consequently, besides face recognition, related works comprise studies and papers dealing with different interconnected topics, typically based on lip-motion representation and analysis for audio-visual speech or speaker recognition [8,10,25,30]. Lip feature extraction from human image is useful in several applications. First systems exploited only audio information. Later, visual parts have been also used, either combined with audio or individually. Mouth regions in lip reading domains can typically be represented in two ways: grayscale pixel-level information or high level visual

Fig. 1. A sequence of frames showing the effects of uttering a sentence on the lower face region. Regions comprised between upper and lower lip, nose-base and upper-lip, chin contour and mouth are all affected to a variable degree depending on both anatomical characteristics and specific speaking habits.

information (geometry, like width, height, surface and mouth opening). In [18] a lip feature extraction algorithm based on Local Binary Patterns (LBP) and Stacked Sparse Autoencoders (SSAE) is presented. According to this method, LBP texture features are extracted from lip images. Then high-level features are extracted using SSAE, which adopts an unsupervised learning to discriminate high-level features. As final step, the method uses fine-tuning in order to improve overall performance. This method features a wide applicability along with high classification accuracy. In [20] the authors propose a spatio-temporal approach to track lip movements, learning from visemes of the French language. It implements three modules. First, a lips tracking system through which lips are segmented using both color and geometric information, since mouth has different color from face skin. Then, a second processing stage implements lip motion tracking by using a particle filter. Finally, visual information are extracted and classified, to allow the recognition of the pronounced viseme.

On a parallel line of research, the work described in [3] can continuously classify if a person is speaking in a video sequence based on lip movement. Firstly, head area is segmented; then, a skin detection technique is applied in order to segment the face area. Next, based on both geometry and color as in [20], the mouth area in each frame is further segmented. A first rough mouth opening detection is based on the fact that the opening area has a darker gray level than its average. Subsequently, only frequency components between 1 Hz to 10 Hz of the detected feature signal are considered to classify the speaking activity by comparing with a threshold. Another method to detect silence sections is proposed in [27]. In this case, the author analyzes geometric parameters such as lip contour's time trajectory, namely interolabial width and height. This method achieved 80% of correct silence detection and 5% of false one. One of the first methods for automated features extraction from lips motion has been proposed in [7] as a potentially valuable resource to improve the resistance of audiovisual authentication systems to replay attacks by means of a liveness-verification

test [22]. Following works [11] and [12], have more formally described the dynamic characteristics of lip-motion which account for its advantage as a secure biometric descriptor. In these approaches the motion component of the captured image sequence is extracted from orientation maps and is then combined to simultaneously extracted speech features to achieve a higher user verification precision. Speaker recognition by lip-motion and speech-reading is also proposed in [5], where a specific experiment based on Hidden Markov Model (HMM) is performed to assess the saliency of lips dynamics. In [26] a statistical approach for lip activity detection and speaker detection in videos is proposed. The main idea is to apply signal detection techniques to a feature extracted from mouth region intensities.

Neural Networks (NNs) have been extensively used in speech recognition as feature extractors in HMM-based speech recognizers [15]. Linear short-term-memory networks (LSTMs) started to replace larger parts of the speech processing system by HMMs. An end-to-end neural network system [13] outperformed HMM-based systems, achieving the best error (16%) on the large Switchboard Hub5'00 speech recognition benchmark [14]. HMM integrated with a multi-boosted learning approach is also exploited by the authors of [17] to devise a comprehensive lip-password enabled speaker verification system. Local spatio-temporal descriptors based on LBP-TOP and a support-vector-machine (SVM) classifier are proposed in [32], while in [19] user authentication through silent utterance of a pass-phrase is approached as a high dimensional time series matching problem. The prospective anti-spoofing advantage of facial dynamics have been first explored in [28] through a combination of Dynamic Mode Decomposition algorithm, LBP and SVM. More recently, a mobile-phone based approach to lip-motion enabled user verification has been proposed in [31] by means of a specialized active shape model algorithm and Gaussian mixture model of lip motion.

With regard to the various descriptors and methods cited above, the proposed approach exploits dynamic facial features not restricted to the sole lips region, but including all the surrounding regions of the lower portion of the face at both the appearance and the motion level, thus resulting in a more discriminant and robust physical/behavioral biometrics, providing a more dependable an accurate recognition performance.

3 Method Description

The rationale of analyzing dynamic local features is motivated by the assumption that the motion patterns of points belonging to those face regions more directly affected during utterance of a given sentence can effectively characterize an individual. In this analysis, we decided to focus only on the visual aspect of speech without considering the audio component, since we are interested in a unimodal system. According to this approach, the succession of frames captured during utterance, contains highly discriminant spatial and temporal information [21]. We are interested not only in the information related to face texture, but also

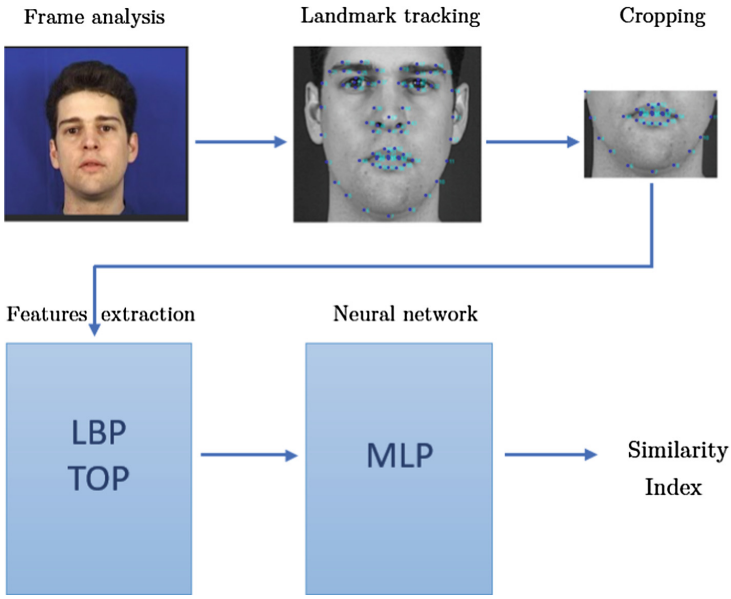

Fig. 2. Schematic view of the overall processing pipelines for the proposed method.

in their changes over time within the frames sequence, producing dynamic textures. As proved by numerous works in the literature [1,2,6,9], indeed, dynamic textures analysis provides the following advantages:

- local texture analysis capturing spatial and temporal information;
- features robust to image transformations;
- computational simplicity;
- good robustness to lighting variation;
- multi-level resolution analysis.

The overall processing pipeline of the proposed method consists of several stages, from subject acquisition to face detection and normalization, and then to dynamic features extraction and recognition, as depicted in Fig. 2. Subject acquisition involves the capture of a video sequence that has to be normalized with regard to the number of frames by means of a re-sampling process aimed at obtaining a clip whose length is consistent to the length of any gallery samples. Each frame of the sequence is therefore analyzed by a face detector [29] that allows to identify the image region in which the subject's face is present. Subsequently, up to 59 facial features are found on the face crop previously detected by means of an efficient landmarks predictor based on [16]. By exploiting these numbered landmarks, the frame is cropped again retaining only the lower face region comprised below the ideal line connecting landmark #2 to landmark #12 (refer to Fig. 3). Finally, the video segments thus obtained are converted into gray-scale and spatially resampled to a resolution of 200×200 pixels. At the

lowest level, the proposed approach is based on the Local Binary Pattern [23], one of the most used and reliable texture descriptor.

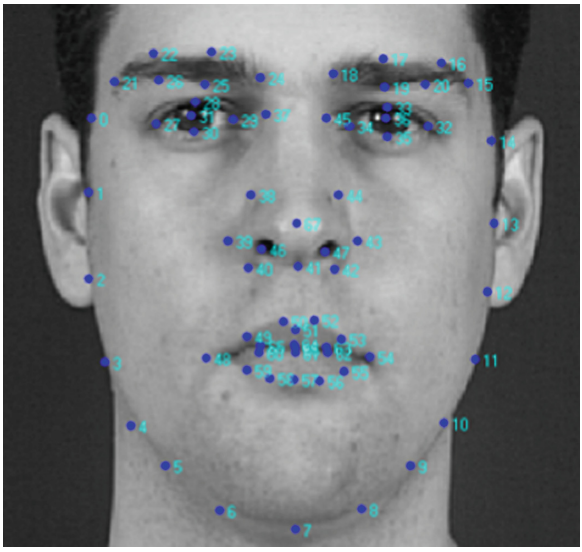

Fig. 3. Facial landmarks considered for facial ROI cropping.

The LBP descriptor replaces the value of each pixel of the image with a decimal value, which is called LBP code and encodes the local structure of the pixel's neighborhood. This is achieved starting from a kernel (central) pixel and considering a serie of neighboring points; for each of them a thresholding is performed with respect to the central pixel value. Concatenating the 0 and 1, calculated through the thresholding operation, a binary value is obtained (see Fig. 4). This value corresponds to the LBP code of that neighborhood. For each block on which LBP is applied, the LBP histogram (i.e. the occurrence of the LBP code in that specific area) is then computed.

The extension to the temporal domain of this simple local descriptor, is known as Volume Local Binary Pattern or VLBP and is particularly suited to describe dynamic-textures such as those resulting by the aforementioned acquisition process. The VLBP descriptor computes the LBP value for each pixel belonging to an area of the space-time volume defined by the dynamic texture, and for each area calculates the histogram, or the occurrence of the LBP codes. To this aim each frame has to be break down into blocks, through a grid. In the present work, we found an adequate partitioning value by using a 4×4 grid applied to the lower face crops resulting by previously described normalization process, resulting in 16 different areas, each of the size of 50 px. The rationale behind this breakdown was preserving the dynamic characteristics for each block. The number of neighboring points that are considered for each pixel

kernel determines the number of bits used to represent the LBP code, therefore the width of the histogram. The latter corresponds to the feature vector relative to the block to which it belongs. The dimensions on which it is applied are X and Y (referring to the spatial domain of each frame) and Z (referring to the frame number in a time sequence) and allow to analyze the local structure not only spatially, but also in its temporal evolution. For each pixel the VLBP code is calculated considering not only the spatial neighborhood, but also the temporal one (see Fig. 5). Consequently, the histogram computed for each block appears to be considerably larger than the histogram resulting from classic LBP, since it takes into consideration more pixels around the central one, leading to a considerable increase in feature vector dimension. To the aim of reducing the computational complexity of the VLBP technique we used a simpler version of it referred as Local Binary Pattern on Three Orthogonal Planes (LBP-TOP), which considers only 3 orthogonal planes for analyzing the local features and is extensively described in [33].

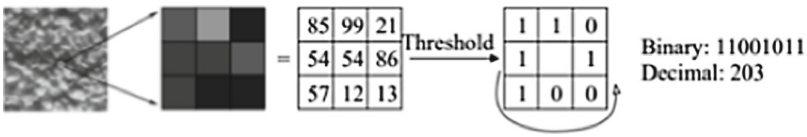

Fig. 4. LBP pattern generation process.

The (LBP-TOP) technique reduces the number of possible patterns by 2(3p + 2) (when considering only 3 planes in the Z dimension) to 3 * 2p, where p represents the number of neighboring points. In this work 36 spaced points were used on a circumference of radius 6, centered on the pixel of interest. The patterns thus obtained are then scaled into integers that can be represented on 8 bits. The binary patterns obtained are extracted from the XY, XZ and YZ planes. The histograms obtained from the three planes are linked together obtaining a single vector of features (Fig. 6). An extension of the original operator is the so-called "uniform pattern", which can be useful to further reduce the length of the feature vector without losing relevant information. Some binary patterns, indeed, occur more often than others in image textures. An LBP code is said to be uniform when it contains only binary patterns that have at most two transitions 0–1 or 1–0. The histogram relative to an LBP technique with uniform pattern will have a distinct bin for each uniform pattern, while it will have a single bin for all non-uniform patterns. In the specific case, considering the value of LBP code expressed on 8 pixels (with possible values between 0 and 255), there are 58 different uniform patterns and therefore the final histogram will consist of 59 bins, where the 59th represents the "other" class.

The resulting feature vector was used to train a fully-connected deep feed-forward (DFFN) neural network schematically depicted in Fig. 7. This network architecture has been preferred over the popular Convolutional Neural Network

(CNN) which typically results much more computationally expensive, requiring a better hardware and more time for training. The number of hidden layers was determined experimentally and the final configuration featuring three hidden levels was the most effective and efficient found. The network provides in output a percentage of probability of belonging to each class, for each sample shown in the testing phase. The class with the highest percentage is then selected, without the use of particular thresholds. The choice of parameters, activation functions and architecture was determined on an experimental basis; a number of tests was therefore performed, modifying the combinations of these variables. It is worth noting that fitting too much parameters to a dataset can led to bad performance on real application or challenging tests, different to training ones. The choice of a good feature representation helped us to reduce this risk to a minimum. However, it is practically impossible to make an absolutely generic model because it would need a infinite dataset. To this regard our approach used processing and features representation as most general as possible, with the ReLu activation function chosen for the input layer, the sigmoid activation function for the three hidden layers and the softmax for the output layer. The number of input nodes was set equal to the size of the feature vector (19824), while the number of output node is determined by the number of possible subjects (295).

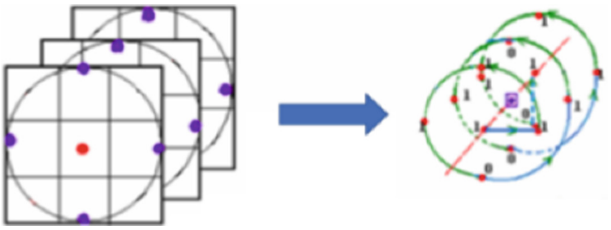

Fig. 5. VLBP descriptor generation.

The network was implemented through the Keras framework with Tensorflow backend; the optimizer and the evaluation metric used are respectively SGD (Stochastic Descending Gradient) and accuracy. All the other parameters of the network, such as the number of epochs, batch size, learning rate, momentum, decay and dropout have been optimized experimentally. The best performing configuration resulted to be the following: epochs = 20, batch-size = 32, learning-rate = 0.1, decay = 0.000001, momentum = 0.

4 Experiments

The experiments described below were conducted the on XM2VTS public database, which is a reference dataset for audiovisual speaker recognition and lip-based speech/speaker recognition. The test-bed was a Fujitsu Celsius machine, featuring an Intel Xeon Octa-Core 2.10 GHz processor and 128 GB of RAM.

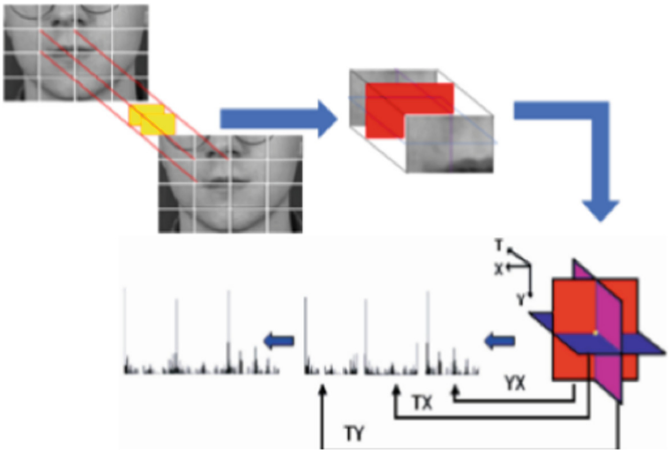

Fig. 6. LBP-TOP descriptor generation.

XM2VTS comprises records of 295 subjects, characterized by a great inter-class variability, both from a demographic and an ethnic point of view. Furthermore, the variability of the same subject in different sessions, such as the growth of beard, presence or absence of the glasses and the change of the hairstyle, also provides wide intra-class variations (see Fig. 8). The dataset, acquired in a controlled environment, is composed of video clips in which each subject pronounces different sentences.

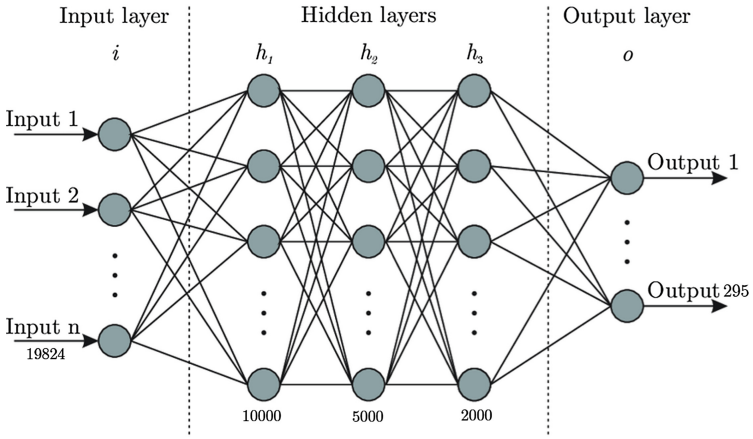

Fig. 7. Network layout of the fully connected Deep Feed-Forward Network architecture used in the proposed method.

More in detail, the dataset is composed of 3 sections: the first contains 4 sessions in which the user pronounces the phrase "Joe took father's green shoe bench out"; the second contains the rotation of the face from left to right (not used in this work); finally in the third there are 4 sessions, for each of which the subjects repeat twice the sequences "zero one two three four five size seven eight nine" and "five zero six nine two eight one three seven four" interspersed with a small pause. The acquisition of 295 subjects has been carried out at a resolution of 720 * 576 at 25 fps.

As mentioned in Sect. 3, several pre-processing operations were required for the use of the video clips. In the first instance, the two 2 sequences contained for each session were separated into the same video file. The separation was achieved by dividing the video in half; although not very refined, this solution has led to optimal results. For each segment thus obtained, the number of frames was calculated, determining the minimum (84 frames) and the maximum (344 frames). Utterance speed is comprehensibly different for each subject according to several factors. One of them depends on the type of sentence to be pronounced, since numbering the digits from zero to nine is simple and natural for everyone as it is mnemonic. For the other sequences, however, the individual needs to learn the order of the digits, and of the words, to then pronounce them quickly. In the latter case, we notice a strong temporal difference in the videos of the different sessions. To this variability is added a further variability due to the characteristic speech speed of every subject. Therefore a resampling operation has been performed in order to uniform the feature vector.

The network was trained on the ordered sequence and three experiments were then conducted to evaluate the robustness of the LBP-TOP descriptor in identifying the subjects. In the first experiment the sequence "zero one two three four five size seven eight nine" was used, dividing it into 80–20% respectively for the train and for the test.

The ROC (Fig. 9) and FAR/FRR (Fig. 10) curves, graphically describe the behavior of the system. The robustness of the proposed approach is confirmed by the EER value of 0.03 and the CMC (Fig. 11), which is 99.8% already at rank-0, reaching 100% at rank-2. This implies a correct classification for almost all the samples, with a very high probability of assignment as shown by the FAR/FRR curve. For the CMC curve only the first 6 of 295 rank have been reported in order to better appreciate the step between rank 0 and 2.

Afterwards, the network was tested on the unordered sequence ("five zero six nine two eight one three seven four") on which she was not trained. In this test the percentage dropped to 98.9%. Finally, to verify the robustness of the model, for the third test, the non-numerical sequence "Joe took fathers green shoe out" was used in the test phase. The accuracy rate was 98.4%. Table 1 report a summary of the results obtained for the three types of sentences pronounced by the subjects in the database. It is worth to note that if a sequence of separated images of the same persons was used rather than a video recording the actual face motion due to speech, the recognition performance would be very low even for a genuine subject. This is directly due to the kind of motion information the

Fig. 8. Inter-class and intra-class variability in XM2VTS dataset.

method is able to represent, which consists in organized changes of the lower face region instead than a generic difference between a sequence of frames.

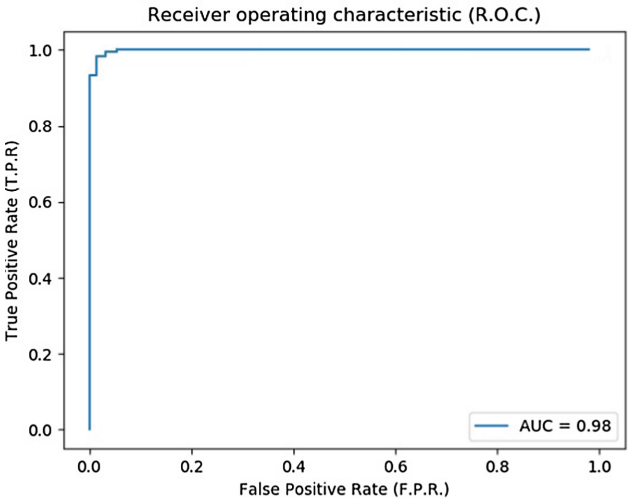

Fig. 9. ROC curve for the proposed method.

The proposed method's behavior, depicted by the results of the experiments, confirms its intrinsic reliability in applicative contexts where the risk of counterfeiting is potentially high. The face dynamic signature provided, indeed, is much more difficult to be forged than any conventional static face descriptor. At the same time, the low Equal Error Rate make a biometric system based on the proposed descriptor suited to medium to high security applications.

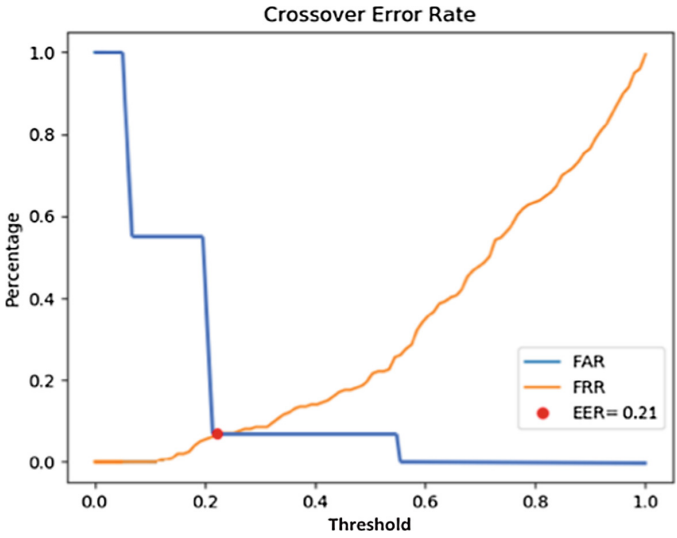

Fig. 10. FAR/FRR curve and EER for the proposed method.

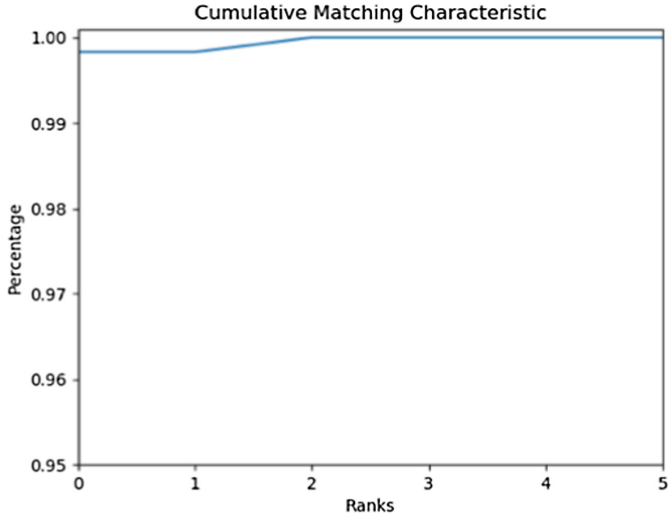

Fig. 11. CMC curve for the proposed method.

Table 1. Resume of the experiments.

Experiment specifications	Ordered numbers	Shuffled numbers	Sentence
Features x block	177		
Sample size	19824		
Train/Test size	1768/589	1768/2355	1768/2357
DNN configuration	19824 – 10000 – 5000 – 2000 – 295		
Accuracy %	99.8	98.9	98.4

5 Conclusions

We presented a method for person recognition exploiting LBP-TOP based representation of dynamic facial features to provide increased dependability in face biometrics thanks to the intrinsic difficulty in forging such a time-dependent descriptor. The proposed deep feed forward network, trained and tested on the audiovisual speech samples from XM2VTS database, delivered a 99.8% recognition rate dropping to 98.4% in challenging testing conditions, achieving in both cases state-of-the-art performance level. Future research will concern more challenging experiments including other public datasets and a direct comparison with the best methods available in literature. An extension of this work could also include the audio component of the speech samples for implementing a bimodal biometric system, to further improve both accuracy and reliability of the proposed method.

Acknowledgments. We gratefully acknowledge the work done by D. Iengo and D. Vanore for implementing and testing the proposed architecture. This work has been partially supported by Italian National Research Project PRIN 2015 (201548C5NT) entitled "*COntactlesS Multibiometric mObile System in the wild: COSMOS*".

References

1. Abate, A.F., Acampora, G., Ricciardi, S.: An interactive virtual guide for the AR based visit of archaeological sites. J. Vis. Lang. Comput. **22**(6), 415–425 (2011). https://doi.org/10.1016/j.jvlc.2011.02.005
2. Abate, A.F., Nappi, M., Narducci, F., Ricciardi, S.: Fast iris recognition on smartphone by means of spatial histograms. In: Cantoni, V., Dimov, D., Tistarelli, M. (eds.) Biometric Authentication BIOMET 2014. LNCS, vol. 8897, pp. 67–74. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13386-7
3. Bandisak, P., Suwansantisuk, W., Kumhom, P.: Classification of speaking activity based on lip features in a sequence of video frames, vol. 11049 (2019). https://doi.org/10.1117/12.2521574

4. Castiglione, A., Raymond Choo, K., Nappi, M., Ricciardi, S.: Context aware ubiquitous biometrics in edge of military things. IEEE Cloud Comput. **4**(6), 16–20 (2017). https://doi.org/10.1109/MCC.2018.1081072
5. Cetingul, H.E., Yemez, Y., Erzin, E., Tekalp, A.M.: Discriminative analysis of lip motion features for speaker identification and speech-reading. IEEE Trans. Image Process. **15**(10), 2879–2891 (2006). https://doi.org/10.1109/TIP.2006.877528
6. Chan, A.B., Vasconcelos, N.: Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE Trans. Pattern Anal. Mach. Intell. **30**(5), 909–926 (2008). https://doi.org/10.1109/TPAMI.2007.70738
7. Chetty, G., Wagner, M.: Automated lip feature extraction for liveness verification in audio-video authentication. In: Proceedings of Image and Vision Computing, pp. 17–22 (2004)
8. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3444–3453, July 2017. https://doi.org/10.1109/CVPR.2017.367
9. Doretto, G., Chiuso, A., Wu, Y.N., Soatto, S.: Dynamic textures. Int. J. Comput. Vis. **51**(2), 91–109 (2003). https://doi.org/10.1023/A:1021669406132
10. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia **2**(3), 141–151 (2000). https://doi.org/10.1109/6046.865479
11. Faraj, M.I., Bigun, J.: Motion features from lip movement for person authentication. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 3, pp. 1059–1062, August 2006. https://doi.org/10.1109/ICPR.2006.814
12. Faraj, M.I., Bigun, J.: Audio-visual person authentication using lip-motion from orientation maps. Pattern Recogn. Lett. **28**(11), 1368–1382 (2007). https://doi.org/10.1016/j.patrec.2007.02.017. Advances on Pattern recognition for speech and audio processing
13. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks, vol. 5, pp. 3771–3779 (2014)
14. Hannun, A.Y., et al.: Deep Speech: Scaling up end-to-end speech recognition. CoRR abs/1412.5567 (2014). http://arxiv.org/abs/1412.5567
15. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012). https://doi.org/10.1109/MSP.2012.2205597
16. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874, June 2014. https://doi.org/10.1109/CVPR.2014.241
17. Liu, X., Cheung, Y.: Learning multi-boosted HMMs for lip-password based speaker verification. IEEE Trans. Inf. Forensics Secur. **9**(2), 233–246 (2014). https://doi.org/10.1109/TIFS.2013.2293025
18. Lu, Y., Gu, K., He, S.: Research on visual speech recognition based on local binary pattern and stacked sparse autoencoder. In: Ahram, T., Karwowski, W., Taiar, R. (eds.) IHSED 2018. AISC, vol. 876, pp. 1082–1087. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-02053-8_165
19. Mendhurwar, K., Mudur, S., Popa, T.: Time series matching for biometric visual passwords. In: ACM SIGGRAPH 2017 Posters SIGGRAPH 2017, pp. 87:1–87:2. ACM, New York (2017). https://doi.org/10.1145/3102163.3102239, https://doi.acm.org/10.1145/3102163.3102239

20. Nainan, S., Kulkarni, V.: Lip tracking using deformable models and geometric approaches. In: Satapathy, S.C., Joshi, A. (eds.) Information and Communication Technology for Intelligent Systems. SIST, vol. 106, pp. 655–663. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1742-2_65
21. Nappi, M., Ricciardi, S., Tistarelli, M.: Deceiving faces: when plastic surgery challenges face recognition. Image Vis. Comput. **54**, 71–82 (2016). https://doi.org/10.1016/j.imavis.2016.08.012
22. Nappi, M., Ricciardi, S., Tistarelli, M.: Context awareness in biometric systems and methods: state of the art and future scenarios. Image Vis. Comput. **76**, 27–37 (2018). https://doi.org/10.1016/j.imavis.2018.05.001
23. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002). https://doi.org/10.1109/TPAMI.2002.1017623
24. Ricciardi, S., et al.: Dependability issues in visual-haptic interfaces. J. Vis. Lang. Comput. **21**(1), 33–40 (2010)
25. Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L.: Static and dynamic 3D facial expression recognition: a comprehensive survey. Image Vis. Comput. **30**(10), 683–697 (2012). https://doi.org/10.1016/j.imavis.2012.06.005
26. Siatras, S., Nikolaidis, N., Krinidis, M., Pitas, I.: Visual lip activity detection and speaker detection using mouth region intensities. IEEE Trans. Circuits Syst. Video Technol. **19**(1), 133–137 (2009). https://doi.org/10.1109/TCSVT.2008.2009262
27. Sodoyer, D., Rivet, B., Girin, L., Schwartz, J.L., Jutten, C.: An analysis of visual speech information applied to voice activity detection, vol. 1, pp. I601–I604 (2006)
28. Tirunagari, S., Poh, N., Windridge, D., Iorliam, A., Suki, N., Ho, A.T.S.: Detection of face spoofing using visual dynamics. IEEE Trans. Inf. Forensics Secur. **10**(4), 762–777 (2015). https://doi.org/10.1109/TIFS.2015.2406533
29. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. CVPR **1**, I511–I518 (2001)
30. Wang, S.L., Liew, A.W.C.: Physiological and behavioral lip biometrics: a comprehensive study of their discriminative power. Pattern Recogn. **45**(9), 3328–3335 (2012). https://doi.org/10.1016/j.patcog.2012.02.016
31. Yuan, Y., Zhao, J., Xi, W., Qian, C., Zhang, X., Wang, Z.: SALM: smartphone-based identity authentication using lip motion characteristics. In: 2017 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 1–8, May 2017. https://doi.org/10.1109/SMARTCOMP.2017.7947043
32. Zhao, G., Barnard, M., Pietikainen, M.: Lipreading with local spatiotemporal descriptors. IEEE Trans. Multimedia **11**(7), 1254–1265 (2009). https://doi.org/10.1109/TMM.2009.2030637
33. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007). https://doi.org/10.1109/TPAMI.2007.1110