DEEP PERSON IDENTIFICATION USING SPATIOTEMPORAL FACIAL MOTION AMPLIFICATION

K. Gkentsidis[1]*, T. Pistola*[1]*, N. Mitianoudis*[1]*, and N. V. Boulgouris*[2]

[1]Democritus University of Thrace
Electrical & Computer Eng. Dept.
Xanthi, Greece

[2]Brunel University London
Electronic & Computer Eng. Dept.
United Kingdom

ABSTRACT

We explore the capabilities of a new biometric trait, which is based on information extracted through facial motion amplification. Unlike traditional facial biometric traits, the new biometric does not require the visibility of facial features, such as the eyes or nose, that are critical in common facial biometric algorithms. In this paper we propose the formation of a spatiotemporal facial blood flow map, constructed using small motion amplification. Experiments show that the proposed approach provides significant discriminatory capacity over different training and testing days and can be potentially used in situations where traditional facial biometrics may not be applicable.

***Index Terms*—** Biometrics, Motion Amplification, Facial Blood Flow

1. INTRODUCTION

Biometrics and their deployment in identification systems is more prevalent than ever before. Most mobile and portable devices have some capacity for biometric identification for the purpose of user access control. Although mobile systems usually rely on traditional biometric traits, such as fingerprints and face, other biometric traits, such as iris and voice, have also evolved significantly over the years and are now widely used in many access control situations.

In parallel to traditional biometric methods, new biometric modalities have emerged, including ear, palm or vein recognition. These new biometrics are deployed either as stand-alone identification modalities or are integrated in multi-modal biometric systems [2]. In essence, each new biometric trait aims at complementing existing biometric solutions in order to improve the performance of traditional biometric systems.

In [3], Wu et al. introduced the concept of video amplification, i.e., a method that can amplify small motions in videos captured by means of an ordinary camera, so that motions can become visible to the human eye. As a first application of the method, the authors of [3, 5] showed that it is possible to amplify and render visible the facial blood flow. An extension of the original framework was presented in [4], where the complex steerable pyramid decomposition was used to apply motion amplification only on the phase component of the decomposition, in order to focus more on the edge information of the video.

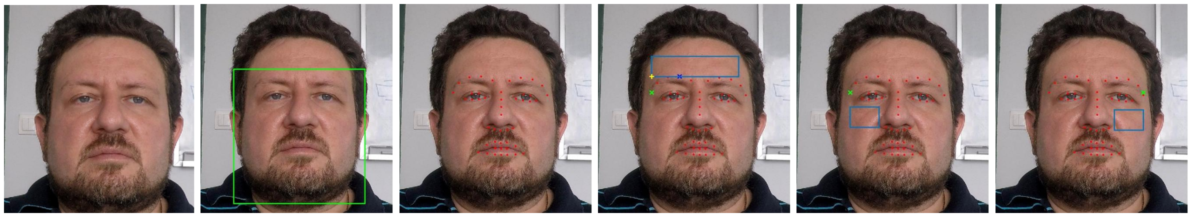
In [1], we proposed a baseline method that uses Facial Blood Flow (FBF) for person identification. FBF is extracted from a person's face using a normal RGB camera, but using motion amplification [3] to enhance and reveal the actual blood flow in common RGB video streams. The proposed method was a contact-less method that did not utilize any traditional facial features. Instead, the method extracts small facial areas, that are not commonly obstructed by facial hair and uses the motion-amplified video to extract spatiotemporal blood flow information, which was shown to work well as a distinctive biometric [1].

In this paper, we improve our baseline method by taking into account the temporal evolution of FBF within a period, which was suppressed in [1]. In addition, a different parallel deep Convolutional Neural Network (CNN) architecture is adopted improving the accuracy of the proposed system over different days.

2. ROBUST FACIAL BLOOD-FLOW BASED IDENTIFICATION

2.1. Video capture and pre-processing

The first stage of the system aims at capturing the subject's face using a camera. The aim is to capture facial blood flow. Thus, we monitor a periodical phenomenon of no more than 2 Hz, since a resting person has an average heart rate of no more than 60-100 pulses per min. This implies that an ordinary RGB camera of 30 fps (i.e., 30 Hz) will suffice for capturing the required facial blood flow, since the Nyquist frequency of the observed phenomenon is 4 Hz. The extra bandwidth provided by the camera can be used to track other frequency content above 2 Hz that may be required to model human identity. We also need to have multiple captures of each subject over multiple days. Some of these will be used for the system's



(a) (b) (c) (d) (e) (f)

Fig. 1. (a) Original face image, (b) Face isolated in a rectangular box using [6], (c) AAM fit on face [8]. The three control points (two from the AAM and another inferred from the other two) are highlighted, (d) Detection of the forehead region using two control points, (e), (f) Detection of the left and right facial regions using the left and right control points respectively.

training, while the rest will be used for the system's testing phase. Finally, natural day light was used during the video capture, in order to avoid additional oscillations by electrical light.

One basic drawback of the motion amplification is that it requires that the subject remains as still as possible. Any extra motion will also be amplified, creating artifacts. Unfortunately, humans can not stay still for more than a few seconds. Therefore, the first stage in our system is to attenuate these visible motions, by using phase-based motion amplification with a negative amplification factor α, in the range of [-1,0).

2.2. Facial Blood Flow Motion Amplification

In this step, we use the Eulerian Video Magnification method by Wu et al. [3] to amplify the facial blood-flow. The amplification factor was set to $\alpha = 120$, while the frequency range of amplification was set between 1 and 2 Hz; this is a good match for the usual range of heart rates (1-2 beats per second) in resting humans. Unlike the work in [3], we are only interested in the amplified blood flow and, therefore, the amplified stream is not added to the original video. In addition, motion amplification is only applied to a grayscale version of the input video to reduce computational complexity. The resultant system configuration yields an image sequence $\mathcal{F}(x, y, t)$ that represents the variation of blood flow on a subject's face.

2.3. Extraction of facial regions of interest

For the extraction of the proposed FBF biometric, we chose three facial areas, namely the lower part of the forehead, the area below the left eye, and the area below the right eye (Fig. 1). These areas do *not* include any facial landmarks (such as eyes, nose, or mouth) and, therefore, are most suitable for the assessment of the discriminatory capacity of the proposed methodology. Further, these areas are usually visible and easy to record.

The forehead region is selected so that it covers the greatest part of the forehead. In most of our experiments, this could be achieved using a rectangular area of 71×201 pixels. The left and right facial areas are both rectangles of dimensions 71×101. To extract the three regions of interest, we first detect the face using the state-of-the-art face detector of Zhu and Ramadan [6]. Subsequently, we identify facial landmarks that facilitate the localization of the three areas of interest. For facial landmark localization in all frames, we use Active Appearance Models (AAM) [7], as implemented in [8]. Once the facial landmarks have been identified (Fig. 1(c)), they are used for inferring the position of the three areas of interest.

The exact positions of the areas below the eyes are determined by correlating key-points on the eyes and the eyebrows. The eyebrows' highest and rightmost points are shown in Fig. 1(d) as blue and green points respectively. Therefore, we can obtain the point, where the forehead starts above the eyebrows. This point's coordinates are the x-coordinate of the eyebrows' highest point and the y-coordinate of the eyebrows' rightmost point. Therefore, we define the starting point (see yellow cross in Fig. 1(d)) of the forehead rectangle. We also define the size (width and height) for the forehead rectangle and we extend the rectangle over the other eyebrow. Using the leftmost eyebrows' point as our reference point (green point in Fig. 1(e)), we move down by 120 pixels in order to find the bottom left corner of the left cheek rectangle. A similar procedure is performed using the rightest eyebrows' point in order to define the area of the right facial area (see Fig. 1(f)). The size of the rectangle and the distance of the rectangle from the highest eyebrow point were chosen based on experimentation. An example of the detected regions of interest is shown in Fig. 1.

3. SPATIOTEMPORAL FACIAL BLOOD FLOW (FBF) TEMPLATES

In [1], where the proposed biometric modality was first introduced, we used facial regions that were smaller and of lower resolution. These regions were joined to form a single rectangular area, which was averaged over time in order to construct a template. That template was constructed by averaging consecutive blood flow images $\mathcal{F}$ representing blood flow during 1 second (i.e., 30 frames). This time period is sufficient to capture at least one blood flow cycle.

Averaging facial blood flow (FBF) images over time is

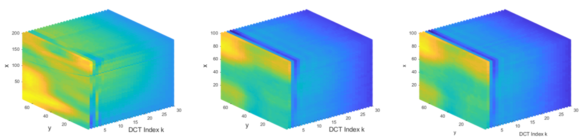


Fig. 2. The FBF DCT features: forehead area, left facial area, right facial area.

computationally efficient, but ignores temporal dynamics, which may possess discriminatory capacity. To address this concern in the present work, we combine FBF image sequences in the form of a spatiotemporal cube from which features are extracted. In this way, the temporal dynamic component of the extracted flow information is retained. In order to avoid having to temporally align FBF image sequences, we apply the Discrete Cosine Transform (DCT), which is sensitive to the presence of different frequencies in a signal and is known to have the capacity for energy compaction. The use of the DCT alleviates problems arising from possible temporal shifts between FBF sequences, which would have an adverse impact on the subsequent training and classification.

In order to calculate DCT features, for each spatial location (x, y) on the facial area of interest, we take the temporal 1D-DCT, which captures blood flow variations over time:

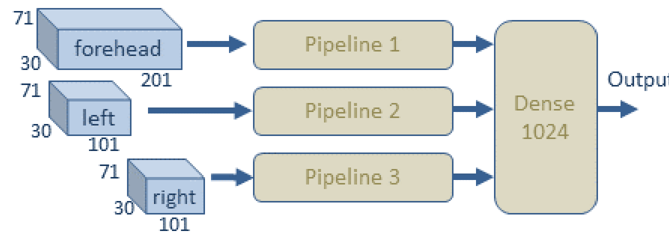
$$\mathcal{T}_F(i,j,k) = \sum_{t=0}^{T-1} \mathcal{F}(i,j,t) \cos[(\frac{\pi}{T}(t+\frac{1}{2}))k]$$

where T is the number of FBF images within one second of recording. The three FBF DCT cubes, which are extracted from the three facial areas of interest, are shown in Fig. 2. These will be input to the classifier.

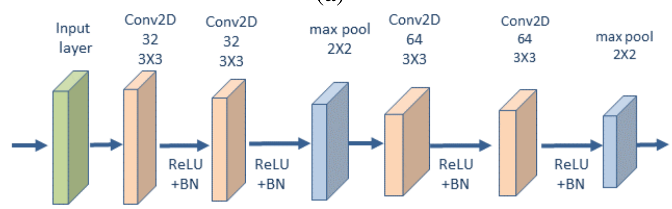
4. PERSON IDENTIFICATION USING DEEP CNN

The three DCT FBF cubes extracted in the preceding stages of the system are input independently to the supervised classifier shown in Fig. 3(a). We use a novel deep architecture that is fundamentally different from the classifier in [1]. As seen, the novel architecture is based on three deep Convolutional Neural Network (CNN) pipelines that process each facial area independently. This resolves problems that may have occurred on the borders between the three cubes had they been joined into one single component.

The CNN pipelines shown in Fig. 3(a) have identical structure, and differ only in the size of the input FBF DCT cube. The prototypical pipeline, shown in Fig. 3(b), consists of four 2D Convolutional Layers. The first layer consists of 32 3×3 filters and uses the ReLU activation function along with a Batch Normalization (BN) stage [10]. The second layer is identical to the first and, in addition, is followed by a 2×2 max pooling layer. The third layer features 64 3×3 filters with



(a)



(b)

Fig. 3. (a) Convolutional Neural Network architecture used with the proposed system, (a) Pipeline for the independent processing of each of the three facial areas.

ReLU and Batch Normalization. The fourth is identical to the third layer and is also followed by 2×2 max pooling. The outputs of the three pipelines are flattened and concatenated, and presented to a dense network of 1024 nodes. The output layer commonly uses the Softmax for classification purposes.

For the training of the proposed architecture, we used the stochastic gradient descent optimizer with the categorical cross-entropy loss function, since we deal with more than two output classes [9]. The learning rate was set to $\eta = 0.01$ with decay 0.01 and a Nesterov momentum of $\mu = 0.9$. The network trained for 60 epochs using a batch size of 64 samples.

5. EXPERIMENTAL EVALUATION

5.1. Dataset

The proposed Facial Blood Flow (FBF) features were evaluated for their effectiveness and robustness as a biometric trait using a new dataset. For the recording of facial image sequences, a GoPro Hero 4 Black camera was used with 1920×1080 resolution and 30 frames per second. A total of 12 subjects were recorded in a room with natural light, in order to avoid oscillations from artificial light sources. The subjects were seated on a chair at a fixed distance from the camera and were instructed to stay as motionless as possible during the recording. This arrangement helped avoid possible scaling and change of posture problems. For each subject, we recorded 18 20-second recordings for training purposes, giving a total 360 seconds per person. For testing, we captured three 20-second recordings from each subject. The recording of training sequences was performed over two different days, while testing sequences were recorded on a different day. This allows us to assess the temporal robustness of the proposed FBF biometric and to reach conclusions on whether the extracted blood circulation patterns remain the same over time. The compiled database is summarized in Table 1.

Table 1. Description of the FBF cross-day dataset. Training sequences were filmed over two different days, while testing sequences were filmed on another day.

Database	
Number of subjects	12
Frame-rate (fps)	30
Resolution	1920×1080
Training sequences/subject	360
Testing sequences/subject	60

5.2. Experiments

We implemented the Movement Motion Attenuation and the Facial Blood Flow Amplification stages based on the code provided by [3] and [4] respectively. The face isolation and AAM fitting stages were based on the method in [8]. We implemented the proposed deep architecture in Python using TensorFlow, Keras and an NVidia Titan X Pascal GPU. We compared the proposed architecture to the one we used in [1], which featured an average image that contains all three regions joined together. The two systems were trained using the training dataset for 60 epochs. We consequently used the testing dataset for assessing the system's performance.

There are significant differences between the system and the experimental protocol presented in the present work and the ones presented in [1]. Specifically, the present work uses features that retain temporal information and, therefore, can lead to conclusions regarding the discriminatory capacity of temporal dynamics in facial blood flow sequences. Furthermore, unlike in [1], where facial image sequences for both training and testing were recorded on the same day, the present work validates the usefulness of the FBF as a biometric by using recordings taken over multiple days. Two different days were used for the training material, while testing material was recorded on a separate day.

Recognition accuracy for the proposed method is presented in Table 2 in comparison to the baseline system in [1]. All systems were tried on the new database. As seen, the architecture in [1], which is based on temporal averages of FBF and therefore ignores temporal dynamics, yields 80% recognition accuracy. The present architecture, however, models temporal dynamics by deploying the DCT along the temporal direction of the FBF feature sequence. The resulting performance improvement provides evidence that temporal information in FBF should be taken into account during classification.

Table 2. Recognition Accuracy for the testing dataset for the previous approach that used the Average FBF template and for the proposed approach with the 3D FBF and the three parallel CNN pipelines.

Architecture	Accuracy
Average FBF Template [1]	80 %
Combined FBF DCT cubes + CNN	81 %
Separate FBF DCT cubes + Parallel CNN	85 %

In addition, processing the three facial areas using three parallel streams also seems to improve the results. In [1], where the three areas were simply stitched together, system performance was impacted by the artificial boundaries between these areas. As seen in Table 2, the single 3D FBF cube, featuring all areas stitched together yields 81% recognition accuracy, while the proposed system with separate parallel pipelines yields 85% accuracy. This highlights the classification architecture that works well with the proposed biometric.

Last, it is most encouraging to observe that, although training and testing sequences were recorded on different days, the proposed FBF DCT cube retains its discriminatory capacity over time. This observation confirms that blood flow patterns on the face remain relatively unchanged over time and can serve as the basis for human identification.

6. CONCLUSIONS

In this paper, we improved the Facial Blood Flow (FBF) biometric proposed in [1]. Our novel feature extraction is based on the modeling of temporal FBF dynamics via the Discrete Cosine Transform. The extracted features are used with an improved deep architecture that processes facial areas of interest independently. Through a new experiment that includes recording on different days, we showed that FBF can be robust over time and has the potential to be used as a reliable biometric in cases where traditional biometrics may not be applicable.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. We would also like to thank the 12 students at the Electrical and Computer Eng. Dep. at Democritus University of Thrace, who participated in our experiment.

7. REFERENCES

[1] T. Pistola, A. Papadopoulos, N. Mitianoudis, N.V. Boulgouris, “Biometric Identification using Facial Motion Amplification”, *IEEE Int. Conf. on Image Processing (ICIP2019)*, Taipei, Taiwan, September, 2019.

[2] A. A. Ross, K. Nandakumar, A. K. Jain, Handbook of Multibiometrics, Springer, 2006.

[3] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W.T. Freeman, “Eulerian Video Magnification for Revealing Subtle Changes in the World”, *ACM Trans. on Graphics*, Vol. 31, No. 4, 2012.

[4] N. Wadhwa, M. Rubinstein, F. Durand, W.T. Freeman, “Phase-based Video Motion Processing”, *ACM Trans. on Graphics*, Vol. 32, No. 4, 2013.

[5] M. Rubinstein, *Analysis and Visualization of Temporal Variations in Video*, PhD Thesis, Massachusetts Institute of Technology, 2013.

[6] X. Zhu, D. Ramanan, “Face detection, pose estimation and landmark localization in the wild”, *Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, USA, 2012.

[7] T.F. Cootes, G.J. Edwards, C.J. Taylor, “Active Appearance Models”, *European Conference on Computer Vision (ECCV)*, Freiburg, Germany, 1998.

[8] A. Bulat, G. Tzimiropoulos, “Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources”, *IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, 2017.

[9] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016.

[10] S. Ioffe, C. Szegedy,“Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”, arXiv:1502.03167.