Secure User Authentication on Smartphones via Sensor and Face Recognition on Short Video Clips

Chiara Galdi*, Michele Nappi, and Jean-Luc Dugelay

EURECOM,
Università degli Studi di Salerno
{chiara.galdi,jean-luc.dugelay}@eurecom.fr,
mnappi@unisa.it

Abstract. Smartphones play a key role in our daily life, they can replace our watch, calendar, and mail box but also our credit card, house keys and in the near future our identity documents. Their increasing use in storing sensitive information, has raised the need to protect users and their data through secure authentication protocols. The main achievement of this work is to make the smartphone not only the cause of problem but also part of the solution. Here, the Sensor Pattern Noise of the smartphone embedded camera and the HOG features of the user's face are combined for a double check of user identity.

Keywords: SOCRatES, Video, SPN, PRNU, Face recognition, HOG features, Smartphone

1 Introduction

An innovative authentication procedure is presented in this paper. On the wake of a previous work, presented by Galdi et al. in [1, 2], that proposes the combination of iris recognition and source camera identification from still images, here the authors present a feasibility study of a system performing face and sensor recognition from a single short video clip. The main objective of this study is to address the need of secure, fast, and easy-to-use authentication systems.

The ever increasing demand of automatic user identification and the exponential spreading of more and more sophisticated smartphones, have led to a natural migration of biometric recognition on mobile devices. However, once that the recognition process is moved to the "end-user side", on one hand it offers the advantage of ubiquitous authentication (the user can potentially get authenticated from any place and at any time) but on the other the anti-spoofing techniques have to be reviewed and revamped in the light of changes in technology. This work proposes an authentication system based on something the user is, namely the face, and something that the user has, that is the personal

* Corresponding author.

smartphone. By observing Fig. 1 it can be noticed that the combination "(5) something the user has + something the user is" assures a higher security level compared to the use of biometrics only (4).

In addition, the increase in security is obtained without making the acquisition procedure too complex. Many works propose multi-modal systems in order to improve recognition performance and robustness to attacks [3], however they often require the user to collect several biometric traits separately. In the latter case, acquisition can take a long time and several sensors can be required for acquiring different biometric traits.

The proposed system, at authentication time, only requires the user to record a short video clip. From that single clip, both face and sensor recognition are performed. This acquisition modality gives room to improve the system by adding for example a liveness detector or by adding a speaker recognition module since during the recording the user could also utter a sentence.

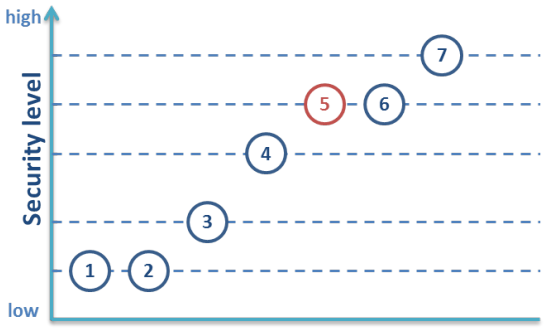

Fig. 1. Authentication systems security levels: (1) Something the user knows; (2) Something the user has; (3) Something the user knows + something the user has; (4) Something the user is or does; (5) Something the user has + something the user is or does; (6) Something the user knows + something the user is or does; (7) Something the user knows + something the user has + something the user is or does.

2 Related works

As mentioned before, the presented work has been proposed on the wake of a research on the subject of the combination of biometry and "hardwaremetry". With the latter we refer to source digital camera identification. The first step in this direction is represented by the work presented by the authors in [1, 2], where iris recognition is combined with sensor recognition, and the input data is a picture portraying the user's eye.

In this paper the input data is a short video clip. It is known that videos are strongly compressed, in particular the ones recorded with smartphones. Thus, in

this case the face has been preferred to the iris, since better performance can be obtained even with low resolution or noisy images. However, the main point that has been investigated is the use of sensor recognition in this context. In particular it has been studied if small video clips, of about 2-5 seconds, are sufficient for the computation of the sensor "fingerprint". Otherwise the enrolment and acquisition processes would be too complex and/or time-consuming and the system would not meet the requirements of speed and ease-of-use.

Concerning sensor recognition, the Sensor Pattern Noise (hereinafter SPN) technique proposed by Lukas et al. [4, 5] and improved by Li [6] has been adopted. The latter achieves optimal performances on still images but sensor identification from videos is much more challenging. The SPN is strongly impacted by video compression, and it is demonstrated that the identification rate can be improved by selecting only I-frames, or a combination on I-frames and first P-frames, for the SPN computation. In [8], Chen et al. propose a technique for determining whatever two video clips came from the same camcorder by mean of the Maximum Likelihood Estimator for estimating the SPN, and of normalized cross-correlation for SPN comparison. In more recent articles, this issue is addressed by selecting only I-frames, or a weighted combination of I-frames and P-frames [7]. Other factors can affect the SPN, for example video stabilization [10] and the additional video compression operated by some website when uploading a video [9].

3 Proposed system

The input of the system consists in a short video clip depicting the user's face. From the recorded video, one single frame would be necessary for the following recognition processes. However, the presence of different images portraying the face could be exploited for performing best template selection. Concerning the sensor recognition module, only the I-frames are suitable, thus two possible strategies could be applied: i) selecting one I-frame that portrays a good face image; ii) selecting two images, the best for each module. Each module, namely the Face and the Sensor recognition modules, process independently the input image. The resulting scores are normalized and combined via weighted sum. In Fig. 2 an overview of the presented system is given.

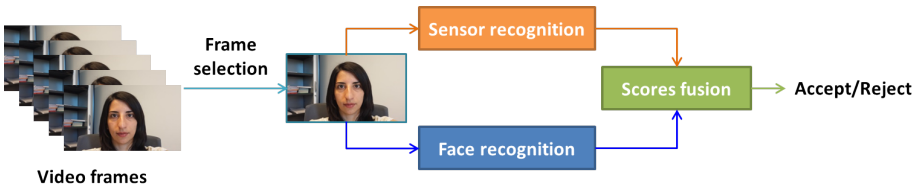

Fig. 2. System architecture.

3.1 Face recognition

The adopted method for face recognition is based on the Histogram of Oriented Gradients (HOG) features [11]. The idea behind this technique is that object appearance and shape can be represented by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. The method consists in dividing the image into small spatial regions, namely "cells". For each cell, a local 1D histogram of gradient directions or edge orientations is computed over the pixels of the cell. The extracted features correspond to the combined histogram entries. For better invariance to illumination and shadowing, contrast-normalization is recommended. This can be done by accumulating a measure of local histogram "energy" over larger spatial regions, namely "blocks", and using the results to normalize all the cells in the block [11]. The resulting HOG descriptors are then used as input of a conventional Super Vector Machine (SVM) based classifier.

3.2 Sensor recognition

Each sensor has a noise pattern due to imperfections during the sensor manufacturing process and different sensitivity of pixels to light due to the inhomogeneity of silicon wafers [6]. This pattern is also reffered as the sensor "fingerprint". Even sensors of the same model can be distinguished by analysing the Sensor Pattern Noise (SPN). This technique has been first presented by Lukáš et al. in [4] and further improved by Li in [6]. The SPN of a sensor is obtained by applying a de-noising filter in the wavelet domain:

$$n = DWT(I) - F(DWT(I))$$

where $DWT()$ is the discrete wavelet transform to be applied on image I and $F()$ is a de-noising function applied in the DWT domain. For $F()$ we used the filter proposed in appendix A of [4].

Since both noise and scene details are located in high frequencies, it is observed that the SPN can be affected by the image content [6]. Li's approach, namely the Enhanced Sensor Pattern Noise (ESPN), is based on the idea that strong SPN components are more likely to have originated from the scene details and thus have to be suppressed, while weak components should be enhanced. The ESPN is computed according to the following formula:

$$n_e(i,j) = \begin{cases} e^{\frac{-0.5n^2(i,j)}{a^2}}, \text{ if } 0 \leq n(i,j) \\ -e^{\frac{-0.5n^2(i,j)}{a^2}}, \text{ otherwise} \end{cases}$$

where n_e is the ESPN, n is the SPN, i and j are the indexes of the components of n and n_e, and α is a parameter that is set to 7, as indicated in [6].

Reference Sensor Pattern Noise For each sensor, the RSPN has been estimated by selecting the I-frames from 9 videos. The videos have a duration that ranges between 2 and 5 seconds. Up to 4 I-frames per video have been extracted - the clips employed for the experiments are very short and thus contain only few I-frames. To extract the RSPN n_r of a sensor, the average SPN over N I-frames is computed:

$$n_r = \frac{1}{N} \times \sum_{k=1}^{N} n_k$$

4 Experimental results

This section describes the dataset employed for the presented experiments and the final results obtained. Performances are assessed in terms of Equal Error Rate (EER), Recognition Rate (RR), Cumulative Match Score curve (CMS), Receiver Operating Characteristic curve (ROC), and Area under ROC curve (AUC).

4.1 Data acquisition and preprocessing

The experiments could be carried out thanks to the publicly available database for SOurce Camera REcognition on Smartphones, namely the SOCRatES database. SOCRatES is currently made up of about 6.200 images and 670 videos captured with 67 different smartphones of 13 different brands and about 40 different models. While images are all of JPEG format, videos are mostly of MP4 file format, with the exception of a few MOV files and one 3gp video.

Videos are treated with simple image processing operations. Video frames are extracted using the *VideoReader* MATLAB function. For the sensor recognition module, I-frames are selected and only a window centered in the image and of size 1024×1024 pixels is employed for the SPN estimation. The images to be given as input to the face feature extractor are first submitted to a face detector module using the *Cascade Object Detector* from MATLAB that implements the Viola-Jones [12] algorithm to detect people's faces. Cropped faces are then resized to 256×256 pixels and converted from RGB to gray level format.

4.2 Sensor recognition performance evaluation

Before presenting the performance of the two modules and of the score fusion, an analysis of the sensor recognition from videos is presented in this section. In particular it has been investigated which I-frame yields to best performances. As mentioned before, SOCRatES contains short videos of about 2-5 seconds. In these short clips only few I-frames are present. The RSPN has been computed over N frames (N = 36 if the videos are long enough to have 4 I-frames each) extracted from 9 out of 10 clips. The 10^{th} video has been used as test sample and its ESPN has been extracted using Li's technique on its I-frames. In the graphs

illustrated in Fig. 3, a comparative performance evaluation shows that using the first I-frame of the test clip, leads to better recognition performances compared to the use of the second or third I-frame. Performances have been assessed on a pool of 630 videos recorded by 63 different sensors.

The performance values are summarized in the following:

- **First I-frame:** EER = 0.3013; RR = 0.3492; AUC = 0.7717.
- **Second I-frame:** EER = 0.3322; RR = 0.2857; AUC = 0.7339.
- **Third I-frame:** EER = 0.3360; RR = 0.2903; AUC = 0.7330.

Compared to the performances reported in [13] obtained on the same dataset but without I-frame selection , an improvement of around 7% of the rate of correct classification has been obtained.

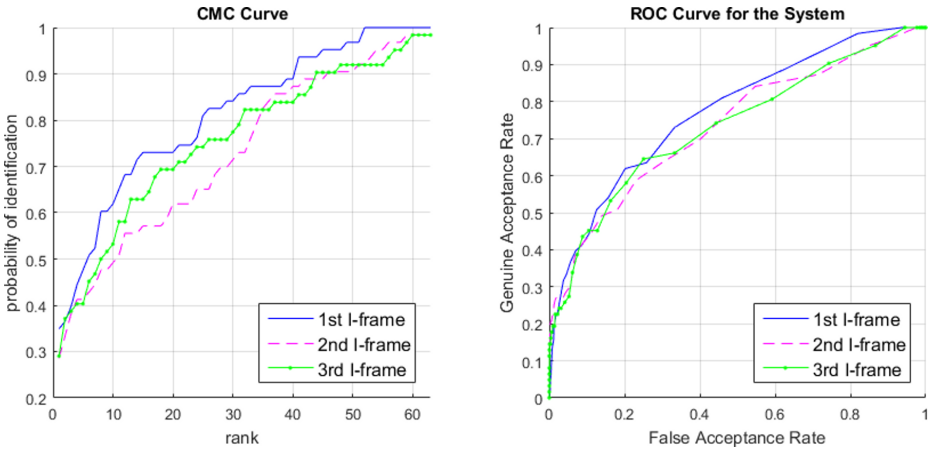

Fig. 3. Sensor recognition performances comparison when using the 1st, 2nd, or 3rd I-frame for SPN extraction.

4.3 System performance evaluation

The videos collected in the SOCRatES database do not portray face images, the system behavior is thus simulated by using the videos as input for the sensor recognition module and pictures of faces for the face recognition. The pictures have been collected with the same devices that recorded the videos. A total of 59 pairs device-face have been then defined. A sample is thus genuine if the combination device-person is enrolled in the system.

The details of the sensor recognition module performances assessment have been already discussed in section 4.2. For what concerns face recognition, 10 face pictures for each user have been collected with the same sensors used to record the videos. The face image are characterized by different pose and illumination.

Eight out of 10 images have been used to train a SVM on the extracted HOG features. One picture, randomly selected from the 2 remaining ones, has been used as test sample.

The performances obtained by the sensor recognition module are: EER = 0.3; RR = 0.35; AUC = 0.77. While the performances obtained by the face recognition module are: EER = 0.07; RR = 0.80; AUC = 0.97. Fusion is performed at score level. The modules contribute equally to the final score since they represent two different entities, i.e. the smartphone and the user, and have the same importance in the computation of the final score. Alternatively, a voting procedure could be used for the final accept/reject decision. The system performances after fusion are: EER = 0.06; RR = 0.83; AUC = 0.97. Fig. 4 reports the performances graph for the aforementioned experiments.

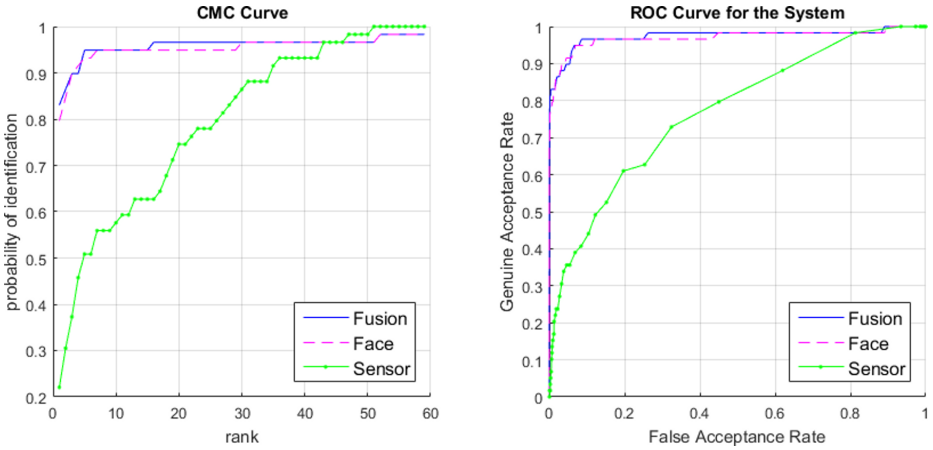

Fig. 4. Single features and fusion performance comparison.

5 Conclusions

An innovative authentication system has been presented. At the best of our knowledge, this is the first work proposing the combination of sensor and face recognition from videos for real-time user authentication. The authors have previously presented a system combining iris and sensor recognition from still images in [1, 2]. Here, the use of videos as input data presents a considerable challenge. In fact, the performances of the sensor recognition module drastically drop when using videos in place of still images (of about the 20% [13, 7]). The SPN is significantly affected by strong video compression. However, by simply selecting I-frames from a set of short video clips, a rate of correct classification equal to 77% is obtained by the sensor recognition module and a rate of 97% is achieved by the combination with the face module.

The objective is to obtain a more secure authentication system by combining different authentication items, namely the user's face and smartphone, while keeping simple and fast the acquisition process. In addition, the proposed acquisition, i.e. a short recording of the user's face, opens the way to further combination with other biometric traits, such us voice, or anti-spoofing or liveness detectors.

The method is tested on a large database of videos collected with 63 different smartphones, namely the SOCRatES database.

References

1. Galdi, C., Nappi, M., Dugelay, J.-L.: Multimodal authentication on smartphones: Combining iris and sensor recognition for a double check of user identity. Pattern Recognition Letters, 0167-8655, 2015. doi: 10.1016/j.patrec.2015.09.009
2. Galdi, C., Nappi, M., Dugelay, J.-L.: Combining Hardwaremetry and Biometry for Human Authentication via Smartphones. International Conference on Image Analysis and Processing, 406-416, Springer International Publishing, 2015.
3. De Marsico, M., Galdi, C., Nappi, M., Riccio, D.: Firme: Face and iris recognition for mobile engagement. Image and Vision Computing, 32(12), 1161-1172, 2014.
4. Lukáš, J., Fridrich, J., Goljan, M.: Digital camera identification from sensor pattern noise. IEEE Trans. Inf. Forensics Security, vol. 1, no. 2, pp. 205-214, June 2006.
5. Goljan, M., Fridrich, J., Filler, T.: Large Scale Test of Sensor Fingerprint Camera Identification. In N.D. Memon and E.J. Delp and P.W. Wong and J. Dittmann, editors, Proc. of SPIE, Electronic Imaging, Media Forensics and Security XI, volume 7254, pages 0I-01 - 0I-12, January 2009.
6. Li, C. T.: Source camera identification using enhanced sensor pattern noise. IEEE Trans-actions on Information Forensics and Security 5(2): pp. 280-287, 2010.
7. Chuang, W. H., Su, H., Wu, M.: Exploring compression effects for improved source camera identification using strongly compressed video. 2011 18th IEEE International Conference on Image Processing, Brussels, 2011, pp. 1953-1956. doi: 10.1109/ICIP.2011.6115855
8. Chen, M., Fridrich, J., Goljan, M., Lukáš, J.: Source digital camcorder identification using sensor photo response non-uniformity. In Electronic Imaging 2007 (pp. 65051G-65051G). International Society for Optics and Photonics.
9. Van Houten, W., Geradts, Z.: Using sensor noise to identify low resolution compressed videos from youtube. In International Workshop on Computational Forensics (pp. 104-115). Springer Berlin Heidelberg.
10. Taspinar, S., Mohanty, M., Memon, N.: Source camera attribution using stabilized video. 2016 IEEE International Workshop on Information Forensics and Security (WIFS), Abu Dhabi, 2016, pp. 1-6. doi: 10.1109/WIFS.2016.7823918
11. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1 (June 2005), pp. 886893.
12. Viola, P., Michael, J. J.: Rapid Object Detection using a Boosted Cascade of Simple Features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. Volume: 1, pp.511518.
13. Galdi, C., Hartung, F., Dugelay, J.-L.: Videos versus still images: Asymmetric sensor pattern noise comparison on mobile phones. In Electronic Imaging 2017