# Full Video Processing for Mobile Audio-visual Identity Verification

Alexander Usoltsev, Khemiri Houssemeddine and Dijana Petrovska-Delacrétaz
*Télécom SudParis, SAMOVAR CNRS, Université Paris-Saclay, 9 rue Charles Fourier, 91011 Evry Cedex, France*

Keywords: Biometrics, Full Video Processing, Face, Speech, Score Fusion.

Abstract: This paper describes a bi-modal biometric verification system based on voice and face modalities, which takes advantage of the full video processing instead of using still-images. The bi-modal system is evaluated on the MOBIO corpus and results show a relative improvement of performance by nearly 10% when the whole video is used. The fusion between face and speaker verification systems, using linear logistic regression weights, gives a relative improvement of performance that varies between 30% and 60% comparing to the best uni-modal system. Proof-of-concept iPad application is developed based on the proposed bi-modal system.

## 1 INTRODUCTION

Mobile devices (like tablets and smartphones) have become an important part of our daily lives, and large amount of personal data is collected and stored in the device's memory. Moreover, with the proliferation of the mobile internet, sensitive information (such as social network accounts, emails and bank accounts) became easier to access. Most of these services require personal authentication. Since mobile devices incorporate both microphone and camera, it is possible to apply a bi-modal biometrics approach based on face and speech modalities to provide protection against unauthorized access.

Mobile biometric system should be fast and autonomous in a way that the user provides only initial media recording (still-image, audio and/or video), and all the processing steps (for example, finding face region and detecting eyes position) should be performed automatically.

However, finding face region and the positions of the eyes in automatic way could be a tricky task due to a wide range of recording conditions (variations of illumination, pose, image quality, etc.). Because modern mobile devices can capture not only still image of the face, but record a video, it is possible to exploit recorded video frames and improve face verification results in hard mobile environment conditions.

On the other hand, if we will pass all video frames to the face verification system on mobile device, it will take long time and consume too much energy to perform computationally demanding biometric verification operations. Therefore, a trade-off between processing the full video and keeping the calculations under a reasonable time should be found.

In this paper, a bi-modal biometric verification system based on voice and face modalities is proposed. The speaker verification system is based on an UBM-GMM (Universal Background Model – Gaussian Mixture Model) method (Reynolds et al., 2000), while the face verification system is using Gabor features and Linear Discriminant Analysis (LDA) modeling (Petrovska-Delacrétaz et al., 2009). For video containing face and speech samples, each modality is first processed separately, and then an overall verification score is computed using score fusion.

For our knowledge, previous evaluation of biometric systems suitable for mobile device used the hand-labeled eye positions, while biometric systems might work different if the face and eye positions are detected automatically rather than annotated by a human (Khoury et al., 2013).

That is why in proposed system the face modality is processed in a fully automated way, including automatic eyes position detection, which involves a whole video sequence processing while keeping balance between calculations time and biometrics system performance.

The proposed system is evaluated on the MOBIO database (McCool et al., 2012). This database, as well as results on face, voice and multimodal fusion,

Usoltsev, A., Petrovska-Delacrétaz, D. and Houssemeddine, K.
Full Video Processing for Mobile Audio-Visual Identity Verification.
DOI: 10.5220/0005667305520557
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 552-557
ISBN: 978-989-758-173-1

is publicly available. Moreover, MOBIO database was recorded on mobile devices, which provides a good estimation of challenges of mobile environment. For every video, database providers extract one still image and manual eyes positions are given only for those still images. This allows us to compare the performance of face verification system when face and eyes detection is performed automatically against using the manual positions of the eyes.

Using the obtained results we develop an iPad proof-of-concept bi-modal biometric application. Despite being early prototype, this application processes full video and performs bi-modal biometric verification in a time of 2-3s (with input video's length of 10s).

This paper is organized as follows. In Section 2 the face and speaker verification systems are outlined. In Section 3, full video processing and score fusion approaches are described. Experimental database and results are exposed in Section 4. Conclusions and perspectives are given in Section 5.

# 2 BI-MODAL VERIFICATION SYSTEM

In this section, baseline face and speaker verification systems are summarized. Two versions of the face verification system are used. In the first version, we suppose that the positions of the eyes are given. In the second version, the positions of the eyes are automatically detected using the Combined Active Shape Model (CASM) (Zhou et al., 2009).

## 2.1 Face Verification

For the automatic version, the baseline face verification system uses Viola-Jones face detector with Haar cascades. The automatic 2D facial landmark location is applied to face region to detect 58 facial landmarks, and from those landmarks the positions of the eyes are extracted.

After geometric and illumination normalization, global facial descriptors are extracted. It involves anisotropic smoothing preprocessing, Gabor features, and LDA face space representation. In this way the discriminative capabilities of LDA systems, that maximize the inter-class variability, and minimize the intra-class distances, are exploited. Gabor filters are applied and only magnitude values are kept to create a vector by concatenation. This vector is projected in the LDA space to reduce its dimension. The size of reduced vector depends on the number of subjects used to create the LDA space. Cosine distance between two reduced dimension feature vectors is chosen as a score to decide whether to accept or to reject the claimant.

## 2.2 Speaker Verification

Well-known GMM-UBM speaker verification system is proposed as a baseline system. The UBM (Universal Background Model) is a GMM with 512 Gaussians trained on a large number of speakers to represent the distribution of the extracted features for a representative population. The hypothesized (or target) speaker model is obtained by adapting parameters of the UBM model using the speaker training speech and a Maximum A Posteriori (MAP) adaptation. In the decision making step, a log-likelihood ratio for a test sequence of feature vectors X is computed as follows:

$$\mathrm{LogLR}(X,\mathrm{target})=\log(p(X|\lambda_{\mathrm{target}}))-\log(p(X|\lambda_{\mathrm{UBM}})) \quad (1)$$

where $\lambda_{\mathrm{target}}$ and $\lambda_{\mathrm{UBM}}$ are respectively the target and UBM models.

The speaker adaptation step creates a model for each of the speakers' voices using samples of their speech. The speaker model once trained allows performing the recognition step by scoring the test speech against the model of the claimed speaker and the UBM model. This score is used to decide whether to accept or reject the claimant.

## 2.3 Score Fusion

One possible way to combine two biometric systems into a bi-modal identity verification system is the score fusion. To follow this approach, scores are first derived using the uni-modal systems. Then, a min-max normalization is performed to produce scores between 0 and 1, where the minimum and maximum score are computed on a development set. Finally, fusion is performed by a weighted sum of the two scores. Linear Logistic Regression (LLR) is exploited to train the optimal weights on a development set.

Combining face and speaker verification systems presents a good solution to overcome the problem, when the face verification system fails to detect faces. In such a case, the fusion system will assign a weight of 1 to the speaker verification and 0 to the face verification instead of a LLR weights. This could be considered as a fusion with a simple quality factor.

# 3 FULLY AUTOMATED SYSTEM

## 3.1 Full Video Processing

For the proposed fully automated system face region and eyes position are found automatically without manual intervention. This system is built on top of the face and speaker verification baseline biometric systems. Speech biometric system works with all audio tracks from videos, however the face biometric system uses only still-images of face. To meet challenges of mobile environment, a full video processing is performed. On the other side, mobile device performance does not allow passing all video frames to face biometric system and user interaction time will be too long. The main idea is to balance between robustness and speed by applying some fast operations to all video frames, such as illumination correction and face region detection, and then choose subset of frames which passed some basic quality measures (e.g. face was found, minimal face dimension is more than 15% of frame dimension). After that, only this subset of frames is passed to more computationally demanding face biometric system.

Steps for the proposed full video processing are:

1) Apply illumination correction filter to all frames of the video. For fast and robust results, logarithmic filter is chosen.
2) Detect face regions on all frames of *log-filtered* video.
3) Select indexes of only three frames where face is detected (two strategies to select frames are discussed below).
4) Apply eye's positions detector to the three selected *filtered* frames.
5) Perform geometric normalization using the eye's positions from step (4) on the three original frames (frame without log filter) according to the indexes selected in step (3).
6) Pass these three frames to face verification system.

Figure 1 shows the scheme of the bi-modal full video processing system. First, two computationally fast operations are applied to all frames from the video: illumination correction with logarithmic filter and Viola-Jones face detector. Then only the biggest detected face in each frame is kept (face size is more than 15% of frame size). After that, computationally more demanding steps of the face verification system are applied on three selected frames.

Two different strategies to select that three frames are evaluated: select three random frames ("random" frames) or frame from the beginning, frame in the middle and frame at the end ("thirds" frames). Influence of selecting "random" or "thirds" frames is reported in section 4.2.2.

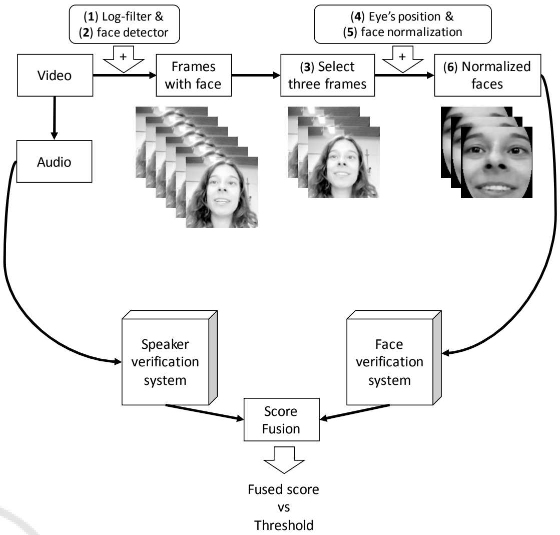

Figure 1: Proposed scheme of full video processing for the bi-modal fully automated biometric system.

# 4 EXPERIMENTS

In this section, the MOBIO database and protocol are described. Then the results related to the face verification system using automatic detection of the face and to its fusion with the speaker verification are exposed.

## 4.1 Database and Experimental Setups

The MOBIO database (McCool et al., 2012) is a bimodal (face/speaker) database recorded from 152 people. The database has a female-male ratio of nearly 1:2 (100 males and 52 females) and was collected from August 2008 to July 2010 in six different sites from five different countries. In total 12 sessions were captured for each individual. The database was recorded using two types of mobile devices: mobile phones (NOKIA N93i) and laptop computers (standard 2008 MacBook). In this evaluation, only the mobile phone data are used, with a sampling rate of 16kHz. Note that a manual annotation of eyes position for the only one image per video data is given. More details about the MOBIO database can be found in (McCool et al., 2012).

Table 1: Number of files (original still images and log-filtered videos) where the automated system fails to detect face in male and female MOBIO development and evaluation sets.

| | | Female | | Male | |
|---|---|---|---|---|---|
| | | DEV | EVAL | DEV | EVAL |
| Enrollment | Still images | 0 | 6 | 2 | 17 |
| | Full video | 0 | 0 | 0 | 0 |
| Test | Still images | 38 | 50 | 30 | 0 |
| | Full video | 2 | 1 | 6 | 0 |

Table 2: EER and HTER computed on MOBIO still images from the development and evaluation partition for the baseline uni-modal systems and their fusion with LLR weight (using ***manual positions of the eyes***).

| | Female | | Male | |
|---|---|---|---|---|
| | DEV (EER) | EVAL (HTER) | DEV (ERR) | EVAL (HTER) |
| Face verification | 8.09 | 13.55 | 7.97 | 8.00 |
| Speaker verification | 11.42 | 11.62 | 10.18 | 9.09 |
| Bi-modal system | 3.27 | 6.14 | 2.73 | 2.86 |

Based on gender of the claimant, two different evaluation protocols for male and female were generated. In order to have an unbiased evaluation, the claimants are split up into three different sets: training, development and evaluation sets:

- *Training set*: The data of this set is used to learn the background parameters of the algorithm (UBM, subspaces, etc.).
- *Development set*: The data of this set is used to tune meta-parameters of the algorithm (e.g. number of Gaussians, dimension of the subspaces, etc.). For the enrollment of a client model, five videos of the claimant are provided. The remaining video files of the claimant serve as testing files.
- *Evaluation set*: The data of this set is used for computing the final evaluation performance. It has a structure similar to the development set.

For the speaker verification system, the feature vector is composed of 20 MFCC coefficients (32 Mel filter bank) together with their first derivatives and the delta energy. This is intended to better exploit the 16KHz range of the MOBIO database. Feature warping and energy-based voice activity detection are performed. The UBM model is trained with 512 Gaussians. SPro 4.1 (Gravier, 2009) and ALIZE 2.0 (Bonastre et al., 2008) software are used to develop the proposed system.

For the face verification baseline system, Gabor filters are applied with 5 scales and 8 orientations convoluted with the face images. An anisotropic smoothing pre-processing is performed and the distance between the eyes is fixed to 50. The SudFrog software (Petrovska-Delacrétaz et al., 2009) is used to develop the system.

## 4.2 Results

To measure the performance of the proposed biometric systems, two different criteria are used: The Half Total Error Rate (HTER) (Khoury et al., 2013) and the Equal Error Rate (EER). To compute the HTER, a threshold θ is defined on the development partition (DEV) at the EER point. This threshold is applied to the evaluation partition (EVAL) to obtain the HTER as follows:

$$\text{HTER} = (\text{FAR}(\theta,\text{EVAL})+\text{FRR}(\theta,\text{EVAL}))/2, \quad (2)$$

Table 3: EER and HTER computed, respectively on the development and evaluation partition, for the face (if located) from still-image and whole video MOBIO database (using ***automatic detection of eyes position***) and speaker verification system, and their fusion using LLR weights (fusion-LLR-quality), and the performance of the fused fully automated system proposed by mccool-icme-2012 (McCool et al., 2012).

| | | Female | | Male | |
|---|---|---|---|---|---|
| | | DEV (EER) | EVAL (HTER) | DEV (EER) | EVAL (HTER) |
| Speaker verification | | 11.42 | 11.62 | 10.18 | 9.09 |
| Face verification, automated | Still images | 13.34 | 15.37 | 12.83 | 9.66 |
| | Full video, Random frames | 11.24 | 15.03 | ***10.59*** | ***8.82*** |
| | Full video, Thirds frames | ***11.08*** | ***14.58*** | 11.31 | 9.16 |
| Bi-modal system | Still images | 5.69 | 7.97 | 5.16 | 4.82 |
| | Full video, Random frames | 5.10 | ***7.92*** | ***4.62*** | ***3.53*** |
| | Full video, Thirds frames | ***5.01*** | 8.40 | 4.86 | 3.69 |
| mccool-icme-2012 bi-modal system | | 10.5 | 13.3 | 10.9 | 11.9 |

where FAR is the False Acceptance Rate and FRR is the False Rejection Rate.

Using these criteria this section studies:

- the influence of applying log-filter with face detection on all frames of the video;
- the performance of the face verification system when the eyes positions are provided manually or automatically;
- the selection of frame's subset;
- the fusion between face and speaker verification systems.

### 4.2.1 Influence of Illumination Correction

The proposed bi-modal system takes advantage of fusion between audio and face verification systems, so it is important to ensure that face can be detected from video even if the light conditions are challenging. To estimate the influence of log-filtering, face detector is applied to still-image (without any correction) and to full video (with log-filter illumination correction). Table 1 gives the number of files (still images and full videos) where the automated system fails to detect the face. Note that a video is considered as containing face only when a face is detected in at least three frames.

Problem with failed face detection in the media means that all possible trials, which involve this media, are lost. In that case bi-modal system relies only on speech data and do not profit from the score fusion.

Usage of illumination correction with logarithmic filter reduces number of failed face detections from 25 files to 0 on enrollment medias, and for test medias from 118 to 9 files. Therefore full video processing does not just improve biometric system performance (section 4.2.3), but in mobile application it also reduces the number of situations when the user is asked to re-record video due to illumination problems.

### 4.2.2 Manual Bi-Modal System

In this section, the manual positions of the eyes are used, which means that the face detection and the automatic 2D facial landmark location modules described in section 2.1 are not considered. The projection space for face verification system is learned on MOBIO training set along with FRGC database (Phillips et al., 2005). 50 users are used for each database, where 30 and 20 images per user are respectively taken from MOBIO and FRGC data. In a similar way, UBM model for speaker verification is learned on MOBIO training set and VOXFORGE data (MacLean, 2012), where 255 female (2h) and 264 male (3h50) files are present.

Table 2 shows the EER and HTER computed respectively on the development and evaluation partition for the baseline uni-modal systems, using the manual positions of the eyes. Fusion of both systems with LLR weights gives a relative improvement that varies between 47% and 65% comparing to best uni-modal system.

### 4.2.3 Fully Automated Bi-Modal System

In this section we compare performance of the bi-modal system when it uses still images from MOBIO dataset against fully processed videos. In both cases face region and eyes position are found automatically without any intervention from the user side. Still images are used "as is" without any modification, and videos are processed as described in section 3.1. For full video processing system, two strategies of subset selection are defined: "random" frames and "thirds" frames (frames from the beginning, middle and the end).

The best score of three frames from face verification system is fused with the score provided by the speaker verification system according to the fusion strategy from section 2.3.

The ERR and HTER computed using still-images and the full videos processing with "random" and "thirds" strategies selection for the face verification and the bi-modal systems are shown in Table 3. The bi-modal system is compared to the fully automated system proposed by (McCool et al., 2012).

The fusion between face and speaker verification systems, using LLR weights, gives a relative improvement of performance that varies between 30% and 60% comparing to the best uni-modal system. The full video random frames for the fusion system shows better results on MOBIO evaluation set comparing to "thirds" frames.

## 5 CONCLUSIONS AND PERSPECTIVES

In this paper, we propose a bi-modal biometric verification system based on voice and face modalities. The face verification part is based on LDA modelling while the speaker one is using an UBM-GMM method. Two versions of the face verification system are developed. In the first version, we suppose that only one still-image is used, while in the second version, system processes all video frames. The proposed fully automated system is evaluated on the MOBIO corpus. The results show that the full video processing gives a relative improvement of performance around 10% for the face system. In addition, the fusion between face and speaker systems relatively improves the performance by 30 to 60% comparing to the best uni-modal system. Additionally, full video processing decreases number of media where face could not be found from 25 to 0 files for enrolment data and from 118 to 9 for test data. Potentially, it allows using more medias for score fusion and improves overall system performance.

After verifying our bi-modal system on the publically available mobile data corpus (MOBIO), we develop an iPad prototype bi-modal biometric application (demo video is available at **https://vid.me/wPJk**). This application performs bi-modal biometric user verification in a time of around 3s.

Future works will be dedicated to conduct more experiments with bi-modal verification with full video processing on a mobile device using existing iPad prototype application. Moreover, different quality measures to select frame subset (such as “mouth close” and “eyes open”,) will be tested to improve the verification results.

## ACKNOWLEDGEMENTS

This work is partially supported by the FUI 15 Equip’Age and H2020-DS-02-2014 SpeechXRays projects.

## REFERENCES

Bonastre, J., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G., Evans, N., Fauve, B., and Mason, J. (2008). Alize/spkdet: a state of-the-art open source software for speaker recognition. In *The Speaker and Language Recognition Workshop, Odyssey*.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models their training and application. In *Computer Vision and Image Understanding*, pages 38–59.

Gravier, G. (2009). Spro: Speech signal processing toolkit, release 4.1.

Khoury, E., Vesnicer, B., Franco-Pedroso, J., Violato, RP., Boulkenafet, Z., Mazaira Fernandez, L.M., Diez, M., Kosmala, J., Khemiri, H., Cipr, T., Saedi, R., Gunther, M., Zganec-Gros, J., Zazo Candil R., Simoes F., Bengherabi, M., Alvarez Marquina, A., Penagarikano, M., Abad, A., Boulaymen, M., Schwarz, P., van Leeuwen, D., Gonzalez-Dominguez, J., Uliani Neto, M., Boutellaa, E., Gomez Vilda, P., Varona, A., Petrovska-Delacrétaz, D., Matejka, P., Gonzalez-Rodriguez, J., De Freitas Pereira, T., Harizi, F., Rodriguez-Fuentes, L.J., El Shafey, L., Angeloni, M., Bordel, G., Chollet, G., and Marcel, S., (2013). The 2013 speaker recognition evaluation in mobile environment, In *ICB '13 : The 6th IAPR International Conference on Biometrics.*

Lowe, D. G. (2000). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110.

McCool, C., Marcel, S., Hadid, A., Pietikainen, M., Matejka, P., Cernocky, J., Poh, N., Kittler, J., Larcher, A., Levy, C., Matrouf, D., Bonastre, J.-F., Tresadern, P., and Cootes, T. (2012). Bi-modal person recognition on a mobile phone: Using mobile phone data. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 635–640.

Petrovska-Delacrétaz, D., Chollet, G., and Dorizzi, B. (2009). *Guide to Biometric Reference Systems and Performance Evaluation*. Springer Verlag.

Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(13):19 – 41.

Stegmann, M. B., Ersbll, B. K., and Larsen, R. (2003). Fame a flexible appearance modelling environment. *IEEE Trans. On Medical Imaging*, 22(10):1319–1331–110.

Zhou, D., Petrovska-Delacrétaz, D., and Dorizzi, B. (2009). Automatic landmark location with a combined active shape model. In *International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–7.

MacLean, K., VoxForge (2012). *Ken MacLean. [Online]. Available: http://www.voxforge.org/home.*

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., ... & Worek, W. (2005, June). Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on* (Vol. 1, pp. 947-954). IEEE.