Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation

Sébastien Marcel, Chris Mccool, Pavel Matějka, Timo Ahonen, Jaň Cernock´, Shayok Chakraborty, Vineeth Balasubramanian, Sethuraman Panchanathan, Chi Ho Chan, Josef Kittler, et al.

▶ **To cite this version:**

Sébastien Marcel, Chris Mccool, Pavel Matějka, Timo Ahonen, Jaň Cernock´, et al.. Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation. ICPR 2010 Contests, Aug 2010, Istanbul, Turkey. 10.1007/978-3-642-17711-8_22 . hal-01318429

HAL Id: hal-01318429
https://hal.science/hal-01318429

Submitted on 30 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation

Sébastien Marcel[1], **Chris McCool**[1],
Pavel Matějka[3], **Timo Ahonen**[2], **Jan Černocký**[3],
Shayok Chakraborty[4], Vineeth Balasubramanian[4], Sethuraman Panchanathan[4],
Chi Ho Chan[5], Josef Kittler[5], Norman Poh[5],
Benoît Fauve[6], Ondřej Glembek[3], Oldřich Plchot[3], Zdeněk Jančík[3],
Anthony Larcher[7], Christophe Lévy[7], Driss Matrouf[7], Jean-François Bonastre[7],
Ping-Han Lee[8], Jui-Yu Hung[8], Si-Wei Wu[8], Yi-Ping Hung[8],
Lukáš Machlica[9], John Mason[10],
Sandra Mau[11], Conrad Sanderson[11],
David Monzo[12], Alberto Albiol[12], Antonio Albiol[12],
Hieu Nguyen[13], Bai Li[13], Yan Wang[13],
Matti Niskanen[14], Markus Turtinen[14],
Juan Arturo Nolazco-Flores[15], Leibny Paola Garcia-Perera[15], Roberto Aceves-Lopez[15],
Mauricio Villegas[16], Roberto Paredes[16]

[1]Idiap Research Institute, CH,
[2]University of Oulu, FI,
[3]Brno University of Technology, CZ,
[4]Center for Cognitive Ubiquitous Computing, Arizona State University, USA,
[5]Centre for Vision, Speech and Signal Processing, University of Surrey, UK,
[6]Validsoft Ltd., UK,
[7]University of Avignon, LIA, FR,
[8]National Taiwan University, TW,
[9]University of West Bohemia, CZ,
[10]Swansea University, UK,
[11]NICTA, AU,
[12]iTEAM, Universidad Politecnica de Valencia, ES,
[13]University of Nottingham, UK,
[14]Visidon Ltd, FI,
[15]Tecnologico de Monterrey, MX,
[16]Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, ES

Abstract

This paper will focus on evaluating the performance of uni-modal face and speaker verification techniques in the context of a mobile environment. This mobile environment (data captured on a mobile phone) presents several challenging recording conditions including adverse illumination, noisy background and variable microphone quality.

This data has been captured in the context of the "Mobile Biometry" (MOBIO) European Project focusing on biometric person recognition from portable devices such as mobile phones. This project is focusing on multiple aspects of biometric person recognition, ranging from research to development, and more particularly on robust-to-illumination face localization and verification, speaker verification in noisy environments, bi-modal fusion, unsupervised model adaptation through time and scalability.

The MOBIO project collected a large audio-visual

database of synchronized face and speech samples recorded from mobile phones across multiple sites in Europe. The distribution of a first part of this database is achieved through the organization of this evaluation: the "Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation".

In total there were nine participants to the evaluation who submitted a face recognition system and five participants who submitted speaker recognition systems.

1. Introduction

Face and speaker recognition are both mature fields of research. Face recognition has been explored since the mid 1960's [7]. Speaker recognition by humans has been done since the invention by the first recording devices, but automatic speaker recognition is a topic extensively investigated only since 1970 [12]. However, these two fields have often been considered in isolation to one another as very few joint databases exist.

For speaker recognition there is a regular evaluation organised by National Institute of Standards and Technology (NIST) [1] called the NIST Speaker Recognition Evaluation. NIST has been coordinating SRE since 1996 and since then over 50 research sites have participated in the evaluations. The goal of this evaluation series is to contribute to the direction of research efforts and the calibration of technical capabilities of text independent speaker recognition. The overarching objective of the evaluations has always been to drive the technology forward, to measure the state-of-the-art, and to find the most promising algorithmic approaches.

Although there is no regular face recognition competition, there have been several competitions and evaluations for face recognition. These include those led by academic institutions, such as the 2004 ICPR Face Verification Competition [40], in addition to other major evaluations such as the Face Recognition Grand Challenge [47] organised by NIST for face recognition.

The MOBIO Face and Speaker Verification Evaluation provides the unique opportunity to analyse two mature biometrics side by side in a mobile environment. The mobile environment offers challenging recording conditions including adverse illumination, noisy background and noisy audio data. This evaluation is the first planned of a series of evaluations and so only examines uni-modal face and speaker verification techniques.

[1] *h*ttp://www.nist.gov

2. Face and Speaker Verification

2.1. Face Verification

TO FINALIZE: Timo ?

The face is a biometric that humans use everyday in passports, drivers licences and other identity cards. However, performing automatic face recognition remains a very challenging task.

Many techniques have been proposed to perform face verification ranging from Principal Component Analysis (PCA) [3] and Linear Discriminant Analysis (LDA) [4] through to feature distribution modelling techniques such as Hidden Markov Models (HMMs) [5] and Gaussian Mixture Models (GMMs) [6]. Normally face recognition evaluations have been limited to databases with highly controlled illumination and pose variations.

Normally the face is controlled to be only frontal. At other times the face can be considered in a set range of pose such as 45 degress or 90 degress. However, the natural range of poses and illumination effects are rarely considered in these experiments. So far this has been to the benefit of the face recognition community because without such databases the individual effects of these variations could not have been explored. However, given the current state of face recognition it is of interest what occurs when there are these naturally occuring pose and illumination problems, and so the MOBIO database aims to address some of these questions by capturing a relatively clean database with few constraints.

2.2. Speaker Verification

The standard scheme of speaker verification is based on feature extraction, followed by universal background model (UBM - based on Gaussian Mixture Models - GMM) from which the target speaker model is adapted on speaker enrollment data. UBM and target speaker model produce likelihoods for the test data. These are subtracted in the logarithmic domain, the resulting score is normalized and compared to a threshold to produce "client/impostor" decision [52]. However there have been proposed many other techniques ranging from Support Vector Machines [15], Joint Factor Analysis [31], or other group based on Large Vocabulary Continuous Speech Recognition systems [60], prosodic and other high level based features for speaker verification [59]. The main problem nowadays is coping with inter-session variability which includes communication channel, acoustic environment, state of the speaker (mood/health/stress), as well as language.

3. MOBIO Database and Evaluation Protocol

3.1. The MOBIO Database

The MOBIO database was captured to address several issues in the field of face and speaker recognition. These issues include:

- having consistent data over a period of time to study the problem of model adaptation,
- having video captured in realistic settings with people answering questions or talking and with variable illumination and poses,
- having audio captured on a mobile platform with varying degrees of noise.

The MOBIO database consists of two phases, only one of which was used for this competition. The first phase (PhaseI) of the MOBIO database was captured at six separate sites in five different countries. These sites are at the: University of Manchester (UMAN), University of Surrey (UNIS), Idiap Research Institute (IDIAP), Brno University of Technology (BUT), University of Avignon (LIA) and University of Oulu (UOULU). It includes both native and non-native English speakers (speaking only English).

The database is being acquired primarily on a mobile phone. There were 160 participants in total who were completed six sessions. In each session the participants were asked to answer a set of questions which were classified as : i) set responses, ii) read speech from a paper, and ii) free speech. Each session consisted of 21 questions: 5 set response questions, 1 read speech question and 15 free speech questions. More details can be found below:

1. **Set responses** were given to the user. In total there were five such questions and **fake responses** were supplied to each user. The five questions asked were:

   (a) **What is your name?**

   (b) **What is your address?**

   (c) **What is your birth date?**

   (d) **What is your credit card number?**

   (e) **What is your driver's licence number?**

   and each question took approximately five seconds to answer (although this varies between users).

2. **Read speech** was obtained from each user by supplying the user with three sentences to read. The sentences were the same for each session and is reproduced below.

   "*I have signed the MOBIO consent form and I understand that my biometric data is being captured for a database that might be made publicly available for research purposes.*

   I understand that I am solely responsible for the content of my states and my behaviour.

   I will ensure that when answering a question I do not provide any personal information in response to any question."

3. **Free speech** was obtained from each user by prompting the user with a random question. For five of these questions the user was asked to speak for five seconds and for ten questions the user was asked to speak for ten seconds, this gives a total of fifteen such questions. The user was again asked to not provide personal information and it was even suggested to not answer the question used to prompt them provided they could speak for the required time.

The collected files are all named according to a particular filename structure. The filename structure is as follows:

PersonID_Recording_ShotNum_Conditions-Channel.mp4

where,
PersonID = Gender + Institute + ID
Recording = Session
ShotNum = Speech Type + Shot
Conditions = Environment + Device
Channel = ChannelID

and
Institute: 0=Idiap, 1=Manchester, 2=Surrey, 3=Oulu, 4=Brno, 5=Avignon
Gender: m=Male, f=Female
ID: from 01 to 99 for each site
Session: ID from 01 to 99
Speech Type: p= set response, l= read speech, r= short free speech or f= long free speech
Shot: ID from 01 to 99
Environment: i=Inside, o=Outside
Device: 0=Mobile, 1=Laptop
ChannelID: ID 0 to 9 (0 - first video/audio channel, 1 -

second video/audio channel)

3.2. The MOBIO Evaluation Protocol

The database is split into three distinct sets: one for training, one for development and one for testing. The splitting is that two sites are used in totality for one split, this means that splits are completely separate with no information regarding individuals or the conditions being shared between any of the three sets.

The training data set could be used in any way deemed appropriate and all of the data was available for use, see Table 1. Normally the training set would be used to derive background models, for instance training a world background model or an LDA sub-space.

Training Splits		
Session number	Usage	Data to use
Session 1	Background training	**All data**
Session 2	Background training	**All data**
Session 3	Background training	**All data**
Session 4	Background training	**All data**
Session 5	Background training	**All data**
Session 6	Background training	**All data**

Table 1. Table describing the usage of data for the Training split of the database.

The development data set had to be used to derive a threshold that is then applied to the test data. However, for this competition it was also available to derive fusion parameters if the participants chose to do so. To facilitate the use of the development set, the same protocol for enrolling and testing clients was used in the development and test splits.

The test split was used to derive the final set of scores. No parameters could be derived from this set, with only the enrollment data for each client available for use. To help ensure that this was the case the data was encoded so that the filename gave no clue as to the identity of the user.

The protocol for enrolling and testing were the same for the the development split and the test split. The first session is used to enrol the user but only the five set response questions can be used for enrollment, see Table 2. Testing is then conducted on each individual file for sessions two to six (there are five sessions used for development/testing) and only the free speech questions are used for testing. This leads to five enrollment videos for each user and 75 test client (positive sample) videos for each user (15 from each session). When producing impostor scores all the other clients are used, for instance if in total there were 50 clients then the other 49 clients would perform an impostor attack. For clarity the enrollment procedure and testing procedure are described again below.

- **Enrollment** data consists of the five **set response** recordings from the first session of the particular user.
- **Testing** data comes from the **free speech** recordings from every other session (the other five sessions) of the users, each video is treated as a separate test observation.

Development and Testing Splits		
Session number	Usage	Data to use
Session 1	Enrollment	**Set questions only**
Session 2	Test Scores	**Free speech only**
Session 3	Test Scores	**Free speech only**
Session 4	Test Scores	**Free speech only**
Session 5	Test Scores	**Free speech only**
Session 6	Test Scores	**Free speech only**

Table 2. Table describing the usage of data for the Testing and Development splits of the database.

4. Face Verification Systems

4.1. Idiap research institute (IDIAP)

The Idiap Research Institute submitted two face (video) recognition systems. The two used exactly the same authentication method (using a mixture of Gaussians to model a parts-based topology) and so differed only in the way in which the faces were found in the video sequence (the face detection method). The systems submitted by the Idiap Research Institute served as baseline systems for the face (video) portion of the competition.

4.1.1 Face Detection, Cropping and Normalisation

Two face detection systems were used:

System 1 is referred to as a frontal face detector as it uses only a frontal face detector. This face detector is based on an MCT-based classifier implemented in [54]. The outputs from this classifier were then modelled using a discriminative method.

System 2 is referred to as a multi-view face detector as it uses a set of face detectors for different poses. Each face detector is implemented as an MCT-based classifiers, the outputs from this were then merged using a normal set of heuristics. More details on this can be found in [54].

From a set of detected faces in the video sequence at most five (5) images were used. The images were selected by retaining the detected frames with the highest score from the face detector; essentially treating the score output from the detector as a confidence score. The chosen images were assumed to be frontal and so the eye positions were estimated from the detected facebox, using these eye positions the images were resized so that the eyes were aligned and resized to have 33 pixels between the two eyes. The face images were then cropped to be a 64×80 image and then illumination normalised by applying a histogram equalisation followed by a Gaussian smoothing.

4.1.2 Feature Extraction

The feature extraction process is performed using the discrete Cosine transform (DCT) and a parts based topology. The parts based topology divides the face into a set of blocks which are then considered to be separate observations, from each observation (block) a feature vector is then extracted. In our particular implementation the face was divided into 8×8 blocks which overlapped in the horizontal and vertical directions by four pixels. From each block DCT features were obtained by keeping the 15 lowest frequency coefficients of the DCT []. Delta coefficients were then obtained to replace the first three lowest frequency coefficients [] and then the x and y position of the blocks were added as another feature. This resulted in feature vectors of 20 dimensions from each block, and so from each image there were a total of 221 blocks or observations obtained.

4.1.3 Enrolment

Before enrolling a user we derive a world or background model Ω_{world} to describe what a face looks like in general. This world model is formed using the data from the training set (the features and faces are extracted and chosen using the same procedure described above). This background model was trained to have 500 mixture components and is subsequently used to initialise the enrollment of a new user and for scoring.

A new user is enrolled by performing background model adaptation of GMMs [16]. The new user is enrolled by using mean only adaptation [50] as implemented in [16] (with a factor of 0.5) from the world model Ω_{world}. Thus for client i we obtain a new GMM Ω^i_{client} by adapting the world model Ω_{world} to match the observations of the client; the client data comes from the enrollment set and uses the same face detection and feature extraction procedures described above.

4.1.4 Authentication

Authentication an observation, $\boldsymbol{x}$, is performed by scoring against the claimed client model (Ω^i_{client}) and the world (Ω_{model}) model. The two models, Ω^i_{client} and Ω_{world}, both produce a log-likelihood score which are then combined using the log-likelihood ratio (LLR),

$$h(\boldsymbol{x}) = \ln(p(\boldsymbol{x} \mid \Omega^i_{client})) - \ln(p(\boldsymbol{x} \mid \Omega_{world})), \quad (1)$$

to produce a single score. Using a threshold tau this score is then assigned to be a true access when $h(\boldsymbol{x}) \geq \tau$ and false otherwise.

4.1.5 Discussion of Results

The results obtained for the two face recognition systems (Frontal and Multi-view) are consistent across both the development and test sets. A summary of the HTERs can be found in Table 3 and it can be seen that the system 2 (using the Idiap Multi-view face detection system) performs slightly better than the system 1 (using the Idiap Frontal face detection system). This is probably due to the fact that more faces are detected using the Multi-view face detector and so there are fewer videos with no faces detected, and so they actually have a chance to correctly verify the user.

	Male	Female	Average
System 1	26.22%	26.64%	26.43%
System 2	25.45%	24.39%	24.92%

Table 3. Table presenting the final results on the Test set for the MOBIO Phasel database.

4.2. Instituto Tecnológico de Informática (ITI)

The approach used for the present contest was based on the work in [63, 65] and is similar to the approach adopted for the ICB 2009 face video competition [49]. From the videos, both for enrolment and authentication, a few key frames are selected based on a quality measure, in this case was based on the confidence of a face not-face classifier. During authentication, for each selected frame a score is obtained and the final score is a combination of the scores for each of the frames.

4.2.1 Face Detection, Cropping and Normalisation

In order to avoid the high correlation between consecutive frames of a video, faces were detected every 0.1 seconds. Each detection was performed on the whole image, in other words, there was no tracking involved. For each video, only the first 2.4 seconds or the first 20 frames with a detected face were used, whichever was shorter. The detected faces were cropped using the estimated eye coordinates and resized to 64×64 pixels. Finally, the images were converted to gray-scale.

System 1 used the *haarcascade_frontalface_alt2* detection model that is included with the OpenCV library. After the detection, a nearest neighbor classifier learnt using [64] was employed to refine the scale and tilt of the detected faces. This classifier consisted of 16 prototypes of size 24×24 pixels, half for face and the other half for not-face, projected onto a 16-dimensional discriminative subspace. The confidence of this classifier was also the one used for the selection of frames for recognition.

System 2 used the face detector from the commercial OmniPerception's SDK. For this system, the scale and tilt of the detected faces was not refined. The measure used for selection of frames was the average of the OmniPerception's SDK detection reliability and the confidence of the same nearest neighbor face not-face classifier from the previous system.

4.2.2 Feature Extraction

From each 64×64 face image, in total 784 local features were extracted. Each local feature corresponds to a 9×9 pixel patch extracted at overlapping positions every 2 pixels. Each local feature is histogram equalized and reduced to 32 dimensions using a PCA basis learnt from all of the local features of 159 world set face images selected by detection confidence. For further detail, refer to [63, 65].

4.2.3 Enrolment

For each enrolment video, the four detected faces with highest confidence were selected. Features are extracted from all of the face images of a user and a kd-tree structure is built in order to make the testing phase faster.

For authentication a background model is also required. For this purpose 159 world set face images were used, 3 per subject, selected based on detection confidence. Again, a kd-tree structure is built to speedup the test phase.

4.2.4 Authentication

For authentication, the score for a given input video x against a client c is given by

$$p(c|x) = \sum_{i=1}^{I} w_i \frac{NN_{c,i}}{F} \quad (2)$$

where the subindex i corresponds to one of the I frames with highest detection confidence, F is the number of local features extracted per face image, and $NN_{c,i}$ is the number local features with a nearest neighbor from the user model c when compared to the background model. There is no score normalization involved in this approach.

The only training performed was adjusting the number of frames used to compute the score I, and the choice of the weights w_i. Both of these parameters were chosen to minimize the error in the development set.

System 1 used the 10 frames with highest detection confidence, i.e. $I = 10$, and for the weights $w_i = q_i / \sum_{j=1}^{I} q_j$, where q_i is the face detection confidence of frame i.

System 2 used the 5 frames with highest detection confidence, i.e. $I = 5$, and constant weights $w_i = 1/I$.

4.2.5 Discussion of Results

The results obtained are basically what was being expected. From previous and the current research it has been observed that this approach gives competitive results while also being quite computationally efficient. There is a difference between the results of the development and test sets, although it is not very significant, which is normal since the parameters were not exhaustively tuned to minimize error rate of the development set.

	Male	Female	Average
System 1	23.97%	19.95%	21.96%
System 2	16.92%	17.85%	17.38%

Table 4. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

4.3. NICTA (NICTA)

We used an off-the-shelf face detection algorithm in conjunction with a modified form of the recently pro-

posed Multi-Region Histogram (MRH) face comparison method [56], which has shown relative robustness to variabilities such as illumination and pose, while retaining scalability. MRH can be thought of as a hybrid between Hidden Markov Model and Gaussian Mixture Model (GMM) based systems. A rudimentary attempt was made to extend MRH from still-to-still to video-to-video comparison. Given the size of the MOBIO dataset, this extension had to maintain scalability while taking some advantage of information from multiple frames. Due to time restrictions, this initial attempt does not exploit all the pertinent information provided by image sequences.

System 1 used ...

System 2 used ...

4.3.1 Face Localisation and Size Normalisation

For face localisation, OpenCV's Haar Feature-based Cascade Classifier [67] is used to detect and localise faces in each frame. The faces are then tracked over multiple frames using Continuously Adaptive Mean SHIFT Tracker [10] with colour histograms. Eyes are located within the face using a Haar-based classifier. If no eyes are located, their locations are approximated based on the size of the localised face. The faces are then resized and cropped such that the eyes are at predefined locations with a 32-pixel inter-eye distance. Two faces sizes are used: 96×96 pixels where possible, falling back to 64×64 otherwise.

4.3.2 Signature Generation and Comparison

The MRH approach is motivated by the concept of 'visual words' (originally used in image categorisation) and can briefly described as follows. A given face is divided into several fixed and adjacent regions (e.g. 3×3) that are further divided into small overlapping blocks (with a size of 8×8 pixels). Each block is normalised to have unit variance and is then represented by a DCT-based low-dimensional feature vector. Each feature vector is then represented as a high-dimensional probabilistic histogram. Each entry in the histogram reflects how well a particular feature vector represents each 'visual word', where the dictionary of visual words is in effect a set of prototype feature vectors. For each region, the histograms of the underlying blocks are then averaged. The 'visual dictionary' is a GMM with 1024 components, built from low-dimensional features extracted from training faces.

For faces with a size of 64×64 pixels, there are 9 regions arranged in a 3×3 layout. For faces with a size of 96×96, 4 additional regions are used (for a total of 13), with the extra regions placed on top, bottom, left and right of the original 3×3 layout.

In a still-to-still scenario, two faces are compared through an L_1-norm based distance between corresponding histograms. For video-to-video comparison, the histograms for a given region are first averaged across the available frames, before using the still-to-still approach. The number of frames used in each video sequence is heuristically capped at 32 frames in order to reduce the computational effort. If a person has several video sequences for enrolment, multiple signatures are associated with their gallery profile.

For each probe video, its signature is compared to the signatures in the gallery to give a raw similarity measurement. Each raw measurement is normalised using a set of cohort images from the training set, using the approach described in [56]. If a person has more than one video available in the gallery, the distance of the probe video to each gallery video is calculated, and the minimum distance is taken.

4.3.3 Discussion of Results

Two submissions were provided for the MOBIO challenge. The initial submission used only closely cropped 'inner' faces (i.e. the inner 3×3 regions), which excluded image areas susceptible to disguises, such as the hair and chin. However, since such periphery information can still give some discriminatory information, the updated submission used 4 additional 'outer' face regions.

The use of the outer regions considerably improved the recognition performance of the female set (HTER fell from 24.46 to 20.83 for the normalised results), but not for the male set (HTER remained around 25). Intuitively, this makes sense as females more often have hair surrounding their heads and uniquely identifiable hair styles as compared to men. This finding has implications for the use of gender specific weightings for inner and outer regions, and also suggests that use of specific gender information may improve performance.

The results further show that lower error rates were achieved on the test sets when compared to the development sets. The reason for this result is mainly due to the OpenCV face detection system locating more faces in the test set (with only 2% of videos with no detected faces) compared to the development set (7% of videos without faces). The training set was in between with 5% of videos without any detected faces.

Participation in this challenge has also highlighted

the importance of a fast and robust face localisation method. Our group's research has so far focused on face recognition, rather than localisation. Since the MOBIO challenge is a system evaluation, we used an off-the-shelf open source face detector (from OpenCV) rather than reinventing the wheel. This turned out to be a major weakness on this particular dataset as the face detector seemed challenged by the pose, glasses, and specular reflection prevalent in the hand-held video recordings.

This initial attempt to extend the MRH face recognition method from still-to-still to video-to-video comparison yielded some promising results with minimal modifications. We aimed for scalability while trying to take advantage of video data by averaging the information over several frames to arrive at a single signature per video. While this approach is very scalable, there was a trade-off in discrimination accuracy.

	Male	Female	Average
System 1	25.84%	25.10%	25.47%
System 1 (norm)	25.39%	24.46%	24.92%
System 2	26.17%	21.99%	24.08%
System 2 (norm)	25.43%	20.83%	23.13%

Table 5. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

4.4. Tecnologico de Monterrey, Mexico and Arizona State University, USA (TEC-ASU)

The CUbiC-FVS (CUbiC-Face Verification System) is based on a nearest neighbor approach to addressing this problem. Despite the simplicity, nearest neighbor approaches have shown strong consistency results in the past. The possibility of extending this approach using the kernel trick is another reason why we found this approach promising.

All of the components were coded in Matlab for clarity and ease of inspection.

4.4.1 Face Detection, Cropping and Normalisation

From the training videos in the development set, it was found that the videos were captured under variable illumination conditions. We therefore used histogram equalization to scale the intensity values uniformly prior to feature extraction. A face detection algorithm based on the mean-shift algorithm (similar to [21]) was then used to localize a face in a given frame. This algorithm is based on online selection of features which are locally discriminative, and thus distinguish between the object and the immediate background. The bounding box detecting the face was then resized to 128×128 pixels in all the images so as to make the dimensionality of the data points consistent.

4.4.2 Feature Extraction

We used the block based discrete cosine transform (DCT) to derive facial features (similar to Ekenel *et al*. [24]), since this feature is known to be robust to illumination changes. Each image was subdivided into 8×8 non-overlapping blocks and DCT was applied to each block. The coefficients were ordered according to the zig zag scan pattern. The first coefficient was rejected for illumination normalization and the top 10 from the remaining AC coefficients for each block were selected to form local feature vectors. To further achieve robustness against illumination, each local feature vector was normalized to unit norm. Concatenating the features from each block yielded the global feature vector for the entire image. The original image had a resolution of 128×128 and thus the dimensionality of the extracted feature vector was 2560. Other features such as Local Binary Patterns (LBP) and Scale Invariant Feature Transform (SIFT) were also tried, but were not found to perform as well as the block-based DCT feature.

4.4.3 Enrolment

Each video stream was sliced into images and the automated face detection algorithm (described above) was applied to detect a face in each image. The detected face was captured and returned in a bounding box surrounding the face. If multiple faces were detected, the areas of each of the bounding boxes were computed and only the face corresponding to the largest box area was considered in this work. We will work on removing this limitation in future work. Images from all the training videos of each subject (as described in the protocols of the challenge) were used for enrolment. For each user U_i, all the feature vectors extracted from the respective video stream were assembled into a training matrix M_i, which was used to train the classifier (described in the next subsection).

4.4.4 Authentication

Given a test vector T, the claim k, and the total number of users enrolled, N, we have to decide whether to accept or reject the claim. Our authentication scheme is based on distance computations using a nearest neighbor classifier (similar to Das [22]). We compute two distance measures, D_{true} and D_{imp}, as follows. D_{true}

is computed as the minimum distance of T from the feature vectors of matrix M_k of the claimed identity k, and D_{imp} is computed as the minimum distance of T from the feature vectors of all matrices other than M_k.

$$D_{true} = min(Dist_{k_i}) \quad (3)$$

where $Dist_{k_i} = (T - V_{k_i})^2$, for $i = 1, 2, ..., x$, V_{k_i} being the feature vectors in matrix M_k. Similarly,

$$D_{imp} = min(Dist_{j_i}) \quad (4)$$

where $Dist_{j_i} = (T - V_{j_i})^2$, for $j = 1, 2, ..., N$ and j not equal to k, $i = 1, 2....x$, V_{j_i} being the feature vectors of matrix M_j.

From these two measures, a score is computed as follows:

$$R = \frac{D_{true}}{D_{imp}} \quad (5)$$

If all the test users are enrolled in the system, then R can be shown to be less than 1 for a client and greater than 1 for an imposter. Thus, the value of R can be used to decide whether the claim has to be accepted or not. The scores were scaled so that clients have a positive score and imposters have a negative score.

4.4.5 Discussion of Results

In the development phase, there were 27 male subjects and 20 female subjects. This resulted in a total of 2025 test videos for males and 1500 test videos for females. Each test video was verified against all possible claims. Our algorithm yielded an EER of 38.62 for males and 41.53 for females on the development data. In the test phase, there were 39 male users and and 22 female users, resulting in 2925 test videos for males and 1650 test videos for females. Our algorithm yielded an EER of 31.36 for males and 29.07 for females, on this test set.

In future work, we plan to extend this approach using kernel functions, and study the performance of different kernel-based feature spaces for video-based face verification.

	Male	Female	Average
System 1	31.36%	29.08%	30.22%

Table 6. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

4.5. InstitutionName (UNIS)

Two algorithms have been tested using the competition protocol. The first system, (UNIS Video 1 and UNIS Video 1(Update)), based on the linear discriminant analysis of Multiscale Local Binary Pattern Histogram (MLBPH) []. The second system, (UNIS Video 2 and UNIS Video 2(Update)), is a heterogeneous, feature level fusion-based system combining MLBPH and Multiscale Local Phase Quantisation Histogram (MLPQH) []. In order to eliminates the score variation caused by condition changes, test-normalisation (T-norm) is applied to both systems.

System 1 used ...

System 2 used ...

System 3 used ...

System 4 used ...

4.5.1 Face Detection, Cropping and Normalisation

In each video, face images are detected by the OmniPerception face detector. The detected face is then aligned geometrically and normalised photometrically by the Preprocessing sequence approach (PS) [62].

4.5.2 Feature Extraction

In our systems, MLBP and MLPQ images are extracted from each of the face image. For MLBP framework[18], Local Binary Pattern operators with 10 different radii from 1 to 10 are applied to the normalised image for multiresolution representation. For MLPQ framework[17], Local Phase Quantisation operators with 8 different sizes are convolved with the normalised image. The resulting pattern images are cropped to the same size and divided into 25 non-overlapping sub-regions. The regional pattern histogram for each scale is then computed. By concatenating these histograms at different scales and then projecting to the Linear Discriminant Analysis(LDA) space, the multiresolution regional discriminative face descriptor is generated for the face matching.

In the updated systems, the XM2VTS frontal face images with Configuration I protocol[41] are used to train the LDA transformation matrix, while the training set of the MOBIO dataset is used to learn the transformation matrix in our basic system.

4.5.3 Enrolment

In the enrolment stage, the updated systems choose the best fifteen frontal face images in each video, based on the confidence of the face detector, and the discriminative face descriptors are extracted as the enrolled feature set. In contrast to the updated system, our basic systems (UNIS Video 1 and 2) only extracts one face image for each video.

4.5.4 Authentication

In each probe video, fifteen face images are chosen in our updated systems while only one face image is chosen in our basic systems. Then the discriminative face descriptors, MLBPH and MLPQH, are extracted for each face image. The similarity measurement of each face image in probe video and each face image of the enrolled subject is obtained by summing the values of a similarity measure, i.e. normalised correlation, of the regional discriminative descriptors together. In order to be robust in the environmental changes, 30 face images of each subject in the provided training set are used as a cohort. Then the maximum similarity score between the enrolled subject and probe face images, and the maximum similarity scores between cohort subjects and probe face images are computed for Test-normalisation(T-norm). The normalised score is used as the final video matching score for UNIS Video 1. In UNIS Video 2, the average of the normalised scores of MLPQH and MLBPH is regarded as the final video matching score. In contrast to the updated systems, only one face image in the video achieving 100% face detector confidence is chosen for video matching. The matching score is regarded as imposter score if the selection requirement is not met.

In order to evaluate the merit of the post-processing, the systems without score normalisation are also reported.

4.5.5 Discussion of Results

The performance of our Updated Systems without score normalisation, (UNIS Video 1 and 2 Update), is significantly better than that of our basic systems. In other words, systems fusing more face image samples can improve the system accuracy, however the computation cost will increase. Therefore, there is a trade-off between the computation cost and accuracy. With the score normalisation, the performance of our systems is improved. One of the reasons for that improvement is that the score normalisation eliminates the score variation caused by condition changes.

	Male	Female	Average
System 1	24.78%	28.03%	26.40%
System 1 (norm)	25.79%	28.67%	27.23%
System 2	25.92%	28.68%	27.30%
System 2 (norm)	27.32%	28.96%	28.14%
System 3	12.04%	14.66%	13.35%
System 3 (norm)	10.35%	13.13%	11.74%
System 4	11.78%	14.04%	12.91%
System 4 (norm)	9.75%	12.07%	10.91%

Table 7. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

4.6. Visidon Ltd (VISIDON)

Visidon face identification and verification system is originally designed for embedded usage, in order to quickly recognize persons in still images using a mobile phone, for example [1]. Thanks to a real-time frame performance, additional information provided by video can be easily utilized to improve the accuracy.

Both object detector (used for face and facial feature detection) and person recognition modules are based on our patented technology. The operation will be covered in the following subsections.

4.6.1 Face Detection, Cropping and Normalisation

Decompressed raw frames were converted into gray scales images, and all operations were performed on these. Subsequent frames do not provide much additional information, and thus we sampled frames in few seconds' intervals only.

A next step after pre-processing was to locate a face in the input frame. For this, we used our own multiview face detector, capable of detecting faces in all orientations starting from 20x20 pixels. If the detector found more than one face per frame, only the most reliable detection was considered. In the case of missed face, the frame was simply skipped.

After locating a face, a geometric correction (similarity transform) was performed to fix the eye locations. To support this, our object detector was run to locate eyes. The face size used for recognition was 80 x 100 pixels. Both face and eye detection were performed on default parameters, without utilizing any temporal tracking. Interesting note for this use case is that most of the faces were acquired from downwards. It is likely that retraining the detectors for this kind of conditions would further improve the detection performance.

Effects of varying illumination were then reduced

from geometrically normalized face images. Inspired by [61], a simple bandpass filtering tuned for typical face and fast processing was used for the purpose.

4.6.2 Feature Extraction

The features are formed utilizing local filters, where each pixel location in a normalized image is associated to a coefficient mask. Using the mask, neighboring pixels affect to the obtained value with predefined weights. This extracts both fine and mid scale structures (depending on the weights and size of the neighborhood) to the feature values extracted. Ignoring largest scales enables recognition of also partially occluded faces. Finally, by extracting statistics of these values, a feature vector of 4608 bytes in length is obtained for one face.

4.6.3 Enrolment

We obtain several candidate faces for one video (one face per each frame considered). As we already skipped most of the frames, these faces now contain more probably complementing information. Here we simply add each successfully processed frame to current individual's codebook, given that maximum amount of images is not exceeded.

4.6.4 Authentication

Input videos are again sampled on few seconds' interval. Each frame under consideration from current video is searched against candidate person data. Measurements from all the processed frames are combined to produce a final probability related value whether the person is who he or she claims to be.

All training of the world model and tuning of the system parameters are done before with data that is independent from the whole MOBIO database. Each comparison is performed independently, as if there were no other persons in a test set or in a query set. No score normalization is performed.

4.6.5 Discussion of Results

Using videos for verification improve the performance compared to still images, although the methods were used in very straightforward manner. The temporal information is limited in using number of frames from one video.

A whole system is designed and implemented as a real-time application running on a mobile phone. All the algorithms are fully optimized and implemented with C language (for portability) using fixed point computation. Running the recognition on a PC is thus very fast, for example, one core of Intel Core2 Duo 2.66GHz processor is capable of handling 100 frames per second when each is compared against 1000 candidates. The fast operation enables also better performance, since more query and prototype faces can be processed in a reasonable time.

Although there were a huge number of frames in MOBIO, the number of individuals in different tests was rather small. For this reason, the results vary between different sets and genders. A failure in enrolling just one individual drops the performance of positive verifications clearly, which can be seen from the figures if the error rate is otherwise low. For example, a development set for females contain 36300 video comparisons, whereas the number of individuals is only 22, and a total failure in enrolling just one of them shifts ROC curve almost 5 percentage units. Difficult individuals have a similar effect on results. Although faces of different persons are not in general much more difficult to recognize - expect against look-alikes - different persons tend to hold their device differently during the verification process. Our recognition method is designed for rather frontal faces, and we are not performing any 3D geometric normalization. Face pointing significantly upwards from the camera causes problems for recognition.

Since the experiments reported here, we have implemented a version that tracks the faces instead of handling these independently.

	Male	Female	Average
System 1	10.30%	14.95%	12.62%

Table 8. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

4.7. University of Nottingham (UON)

Our method bases on 4 different facial descriptors, 2 different subspace learning methods and Radial Basis Function SVM for verification. Four facial descriptors are Raw Image Intensity, Local Binary Patterns, Gabor Filters and Local Gabor Binary Patterns. Two subspace learning methods are Whitened Principal Component Analysis and One Shot Linear Discriminant Analysis. Verification is performed using RBF SVM.

System 1 used ...

System 2 used ...

4.7.1 Face Detection, Cropping and Normalisation

We used OpenCV's Haar Feature-based Cascade Classifier [66] with the following parameters: cvHaarDetectObjects(gray, cascade, storage, scale_factor=1.1, min_neighbor=3, flags=0, min_size=cvSize(150, 150)). Then, PCA is used to learn the face subspace and all regions which are far from that subspace have been removed. Finally, the region containing the largest percent of skin color has been selected as a single face region candidate. Within that region, we detect the eyes and normalize the face so that two eyes are at two specific locations and resize the face to 64×64. The eye locator works as follows. Eye region is defined as the upper half of the face image and eye detection works on the left and right half of the eye region respectively for the left and right eyes. Firstly it detects rotationally symmetric (circular) objects using generalized symmetry transform. Edges are detected using Canny edge detection and all edge points are paired to vote the midpoint of their connection for potential symmetry centers with symmetry scores. The symmetry scores are contributed by the symmetry and magnitude of image gradients at the pair of edge points. An expected size of eyes or irises is also compared with the actual distance between the pair of edge points to scale the score. The original image is therefore transformed to a symmetry map and the point in the map with the maximal symmetry score is selected as the position of eye candidates. Next a circular shape template for iris is used to locate the iris in the neighborhood of eye candidates by an exhaustive search or random search. With properly defined energies based on the edge map, the symmetry map and grayscale values of the original image, the search explores the iris state space to find the state where the energy is minimized. The detector finally outputs the coordinates and size (radius) of the iris.

4.7.2 Feature Extraction

We used 4 different features: Raw Image Intensity (IN), Local Binary Patterns (LBP), Gabor Filters (Gabor), and Local Gabor Binary Patterns (LGBP).

Raw Image Intensity is simply the grey intensity of each pixel. The length of the feature vector is the number of pixels, 4096 (64×64).

LBP was first applied for Face Recognition in [2] with very promising results. In our implementation, the face is divided into non-overlapping 8×8 blocks and LBP histograms are extracted in all blocks to form the feature vector whose length is 3,776 ($59 \times 8 \times 8$).

Gabor Filter with 5 scales and 8 orientations are convoluted at different pixels selected uniformly with the downsampling rate of 4×4. The length of the feature vector is 10,240 ($5 \times 8 \times 16 \times 16$).

The last type of feature is LGBP [70, 58, 29]. There are total of 151,040 ($5 \times 8 \times 59 \times 16 \times 16$) LGBP features. All features are sorted in descending order of their variances. The first 15,000 features are selected to form the feature vector.

4.7.3 Enrolment

Locally Linear Embedding (LLE) [55] is used to select best frames from videos. We apply LLE for all frames to reduce dimension then use K-clustering to select best 5 frames from each video.

4.7.4 Authentication

Whitened PCA (WPCA) and One-shot LDA (OS-LDA) [69] are used to compute the similarity between two input faces. Four features and two subspace methods form a total of 8 similarity scores which can be considered as a 8-D vector. This 8-D vector is passed to RBF SVM for verification.

RBF-SVM parameters (c and γ) are trained using cross validation using LIBSVM library. Training and testing sets are splited so that they don't share any common subject. In other words, any subject appears in either training or testing set, exclusively. If training set and testing sets share common subjects, over-fitting happens as shown in the results. The final score is a number between 0 and 1 which is the probability of two input faces matching.

We don't perform any score normalization method.

4.7.5 Discussion of Results

	Male	Female	Average
System 1	49.21%	48.49%	48.85%
System 2	29.80%	23.89%	26.85%

Table 9. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

4.8. National Taiwan University (NTU)

4.8.1 Face Detection, Cropping and Normalization

The first step of our system detected faces, and an additional step was applied to reject false face detections. The following gives the detailed steps:

1. For every frame, we detected faces using the OPENCV face detection function with a relatively high threshold for the first run. Specifically, a location in a video frame was regarded as a face only if more than 40 face rectangles were returned by the face detection algorithm. If the first run failed to detect any face, then the second run of face detection with a lower threshold was performed. Our system performed at most 3 face detection runs. The thresholds for the 3 runs were 40,20,5 face rectangles, respectively. This step tends to obtain faces of good quality, if possible.

2. We detected at most one face in each video frame. For each face detected by the OPENCV face detection function, we applied the Active Shape Model (ASM) to locate fiducial points on this face. If ASM failed to locate facial points on this face, this face was ignored.

3. Then we performed the geometric normalization of the face image. We first calculated the eye centers, and rotated the face to make the line passing through eye centers horizontal. This step corrects the in-plane rotation.

4. Then we calculated the mouth center, and the (horizontal) distance between eye centers. Assume it equals to x.

5. We also calculated the vertical distance between the center of eyes and mouth center. Assume it equals to y.

6. We defined the face borders:

$$\begin{aligned} d_L &= 0.5x \\ d_R &= 0.5x \\ d_U &= 0.6y \\ d_B &= 0.7y, \end{aligned}$$

   where d_L is the horizontal distance from the left border to right eye, d_R is the horizontal distance from the right border to left eye, d_T is the vertical distance from the upper border to the center of eyes, and d_B is the vertical distance from the lower border to the mouth center.

7. We cropped the face from the image based on the face borders, and resized the cropped face into 80x100 pixels. The ratio between the width and the height typically changes after this resizing. In our experience, this step corrects the out-of-plane rotation to some extend, and it works well when face are under large out-of-plane rotation. Facial images were converted to 8-bit grayscale images. To alleviate the impacts made by illumination variations, all samples were processed to have mean 128 and variance 25.

8. To reduce the false face detection, we employed a Support Vector Machine (SVM) to classify faces and non-faces. We run our system on photos from the World-Wide Web, and collected false face detection examples as the negative examples of the face-nonface SVM.

9. To guarantee that the detected faces were well aligned, an additional PCA-based classifier that classifies a face into a well aligned face and a poorly align face was also employed.

4.8.2 Feature Extraction

We applied the Probabilistic Facial Trait Code (PFTC), which is an extension of our previous work, the Facial Trait Code (FTC) [33] . FTC is a component based approach. It defines the N most discriminative local facial features on human faces. For each local feature, some prominent patterns are defined and symbolized for facial coding. The original version of FTC encodes a facial image into a codeword composed of N integers. Each integer represents a pattern for a local feature. Unlike FTC, The PFTC encodes a facial image into a codeword composed of N probability distributions. These distributions gives more information on similarity and dissimilarity between a local facial image patch and prominent patch patterns, and the PFTC is argued to outperform the original FTC. The associating study is currently under review. In this competition, we used 100 local facial features, each had exactly 100 patterns, and it made up a feature vector of 10000 real numbers for each face.

4.8.3 Enrolment

We collected at most 10 faces (in 10 frames) from an enrollment video. Each collected face was encoded into a gallery codeword using PFTC.

4.8.4 Authentication

We collected at most 5 faces from a testing video. Each collected face was encoded into a probe codeword using PFTC. Then, this probe codeword was matched against known gallery codewords. Assume an enrolled identity has M faces, and a test video contains N faces detected by our system. The distances between all the enrolled

face and test face pairs were calculated, resulting a M-by-N distance matrix. The verification score was the *maximum* score among these $M \cdot N$ scores. We did not perform the gallery normalization on scores of each testing data.

4.8.5 Discussion of Results

It took us three man-months to develop and modify our system for this evaluation. The training data for our algorithm consisted faces collected from the world-set provided by the MOBIO contest, a subset of FERET, a subset of FRGC 2.0, and faces collected in our laboratory using ordinary web cameras. The training data included about 5000 facial images from 500 different identities. The training of our algorithm (PFTC) using these data took about 3 full days on one PC, and it required roughly 1.8GB memory at most.

For enrollment, we collected 10 faces from 10 frames in a video. The two frames in which faces were collected are parted by 10 frames at least. It took roughly a second to enroll a face, so it took roughly 15 seconds for the enrollment of one user. The approximate processing time for the verification of one video file against one user was roughly 0.3 second. This process required 50KB for each face. Assume we collect 5 faces in a testing video, and a user has 50 faces enrolled in the database, then the memory requirement for the verification of one video file against one user is roughly 2.68MB.

It seems that we achieved average performance in this evaluation. Our performance can be improved if we collect more faces from a single video sequence. A video sequence typically includes more than 300 frames, and we only use 10 frames and 5 frames for enrollment and testing respectively. The reason we use only a very small subsets of all available frames is to reduce the complexity, given that we had very limited time before the deadline for the submission of our results.

	Male	Female	Average
System 1	27.98%	36.56%	32.27%
System 1 (norm)	20.50%	27.26%	23.88%

Table 10. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

4.9. iTEAM, Universidad Politecnica Valencia (UPV)

The system proposed by UPV is based on the HOG-EBGM [3] algorithm. This algorithm is used to extract biometric information from the face pixels. The HOG descriptor is a local statistic of the orientations of the image gradients around a facial landmark. Compared to other local features, the HOG descriptors are more robust against changes in illumination, small displacements and small rotations [42]. The HOG descriptors are also used to detect the eyes which is an important step for the face normalization.

To deal with the multiple faces detected in each video our system selects a small set that contains *the best* faces.

4.9.1 Face Detection, Cropping and Normalisation

We used two different off-the-shelf algorithms for face detection in the development and test stages respectively. Initially, we used the OpenCV AdaBoost face detection system [34], however since we found that this algorithm was not able to detect any face in some enroll development videos we changed to a commercial closed solution [44] for the updated release. Although the face detection results provided by the Verilook algorithm are slightly better, we found that the improvement in the recognition results is minimal.

Detected faces are normalized using eye coordinates. To detect the eyes we have developed a two stage algorithm that first detects eye candidates using Haar features and Adaboost, and second a SVM classifier is used to select the best eye-pair using HOG descriptors [43].

Once eyes are detected, the normalization of the face is performed by cropping the face region to a 125×145 image and placing the eyes at fixed locations (coordinates $[25, 35]$ and $[100, 35]$ respectively).

It should be mentioned that in our updated release we introduced also a kalman filter as explained below to track the eyes and reduce the detection noise. The contribution of this step to improve the recognition results was more important than the change of face detection algorithm.

4.9.2 Feature Extraction

Once faces are extracted and normalized in scale and translation, we extract features using our HOG-EBGM algorithm [3]. Our algorithm is similar to the well known Elastic Bunch Graph Matching (EBGM) approach proposed by [68] in which biometric information is extracted at 25 facial landmarks using Gabor features. The key improvement of our approach is

that we replace Gabor features by HOG descriptors. These descriptors are more robust to small displacements and illumination changes. The interested reader can check [42] for a comparison between HOG-EBGM and Gabor-EBGM.

Our HOG descriptors are much like SIFT features [36], except that SIFT features are extracted at the local extrema of a scale-space representation of the image and normalized in scale and rotation. We deliberately skip these two normalization stages because our input faces are already normalized in scale and rotation. However as is the algorithm proposed by Lowe, each HOG descriptor is also a histogram in which the bins form a three dimensional lattice with $N_p = 4$ bins for each spatial direction and $N_o = 8$ bins for the orientation for a total of $N_p^2 N_o = 128$ components. In our work, each spatial bin is a 5×5 pixels square. This size was chosen accordingly to the distance between eyes of the normalized faces.

Finally, the feature vector extracted for each face is the concatenation of all the HOG descriptors obtained at each facial landmark. This results in a the feature vector of $25 \times 128 = 3200$ components.

Since the dimensionality of this feature vector is too high we use Kernel Fisher Analysis (KFA) [35] to perform dimensionality reduction and non-linear feature extraction. The KFA was trained using face images from the FERET database (600 images corresponding to 200 individuals) [48] and ten face images of each person of the MOBIO training set. We made experiments using only the FERET and only the MOBIO training set, but the best results were achieved when these two sets were combined together. This can be explained because the FERET images include a higher number of different people, on the other hand the MOBIO training set can better model the intra-person variability because more images per person are available. The final number of features per face after dimensionality reduction is 140.

4.9.3 Enrollment

To enroll a new person we just select the N faces with highest confidence from the corresponding videos and store the set of feature vectors from each of those faces as a model for the person. We used two different confidence values in the initial and final releases to select the best images from the many detected faces in the videos.

As it is known, almost every face detection system produces a number of hits around each real face which are usually clustered into one detection. In our initial system, we used this number of hits provided by the OpenCV Adaboost face detection system as the face confidence. However, we found that with this confidence measurement we were missing the important information about the goodness of the eyes localization, which in turn is very important to obtain a *good* normalized face. For this reason in our final release we introduced a simple Kalman filter to track the location of the eyes in the video. Then, we use the Euclidean distance between the detected eyes position and the corresponding Kalman predictions as a measurement of the face confidence. This measurement allows to select faces with low head motion (which are sharper) and with small noise in the eye detection stage.

In the development stage we made experiments with different number of faces in each person model. We found that a number of $N = 10$, was a good trade off between complexity and accuracy. In fact, we did not get significant recognition improvements using higher values of N which indicates that a good representation of the person was already obtained with just ten faces.

4.9.4 Authentication

Similar to the enrollment stage, to authenticate a video we first extract its best faces from the query video. We also used the two different face confidence measurements explained above in the initial and final releases to select the best faces among the multiple detections.

Once the dimensionality-reduced feature vectors are extracted for the best test faces using HOG-EGBM and KFA, authentication is performed comparing each of these vectors with those stored for the enrolled person. All pair-wise comparisons are performed using cosine distance and the minimum value is used as the final similarity score between the query video and the person model.

4.9.5 Discussion of Results

The face recognition system provided by UPV achieved good performance on the MOBIO data with a minimal tuning of the recognition algorithm. The only part of the algorithm that was particularly tuned was the KFA feature extraction, in which faces from FERET and MOBIO training dataset were used. This particular tuning gave an improvement of about 2% in the equal error rate using the development data.

The difference in recognition performance between males and females is also statistically insignificant, which is consistent with the fact that we never designed our algorithm to be gender dependent (using hair style features for instance).

We did not observe any significant difference on the recognition results on the development and test sets,

which shows that the difficulty of both datasets was similar and it is also proves that our system is not tuned to any particular dataset.

Finally, we also observe a small improvement in the updated release of our algorithm that is produced by a better selection of good faces using the Kalman filter described above.

	Male	Female	Average
System 1	23.74%	23.70%	23.72%
System 2	21.86%	23.84%	22.85%

Table 11. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

5. Speaker Verification Systems

5.1. Brno University of Technology (BUT)

The Brno University of Technology submitted two audio speaker verification systems and one fusion of these two systems. First system is Joint Factor Analysis and second one iXtractor system. Both systems used for training the MOBIO data but also other data mainly from NIST SRE evaluations.

5.1.1 Voice Activity Detection and Speech Segmentation

Speech/silence segmentation is performed by our Hungarian phone recognizer [57, 39], where all phoneme classes are linked to 'speech' class. We used only speech class for further modeling.

5.1.2 Feature Extraction

We used 24 mel-banks, 25ms window with 10ms shift for computation of 19 MFCC on the audio files sampled at 8000Hz. The features are augmented with energy and with their delta and double delta coefficients, making 60 dimensional feature vector. Features are short-time gaussianized with window of 300 frames (3 sec) [11].

5.1.3 Enrollment

Universal Background model One gender independent and two gender dependent universal background models (UBMs) with 2048 Gaussians were trained on Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, and NIST SRE 2004 and 2005 telephone data. In total, there were 16307 recordings (574 hours) from 1307 female speakers and 13229 recordings (442 hours) from 1011 male speakers.

System 1 - Joint Factor Analysis The Joint factor analysis (JFA) system closely follows the description of "Large Factor Analysis model" in Patrick Kenny's paper [31], with the speaker model represented by mean super-vector: $\mathbf{M} = \mathbf{m} + \mathbf{Vy} + \mathbf{Dz} + \mathbf{Ux}$, where $\mathbf{m}$ is speaker-independent mean super-vector, $\mathbf{U}$ is a subspace with high inter-session/channel variability (eigenchannels) $\mathbf{V}$ is a subspace with high speaker variability (eigenvoices) and $\mathbf{D}$ is a diagonal matrix describing remaining speaker variability not covered by $\mathbf{V}$.

The two gender-dependent UBMs are used to collect zero and first order statistic for training two gender-dependent JFA systems. First 300 eigenvoices are trained on the same data as UBM, although only speakers with more than 8 recordings were considered here. For the estimated eigenvoices, MAP estimates of speaker factors are obtained and fixed for the following training of eigenchannels. A set of 100 eigenchannels is trained on SRE 2005 auxiliary microphone data (1619 and 1322 recordings of 52 females and 45 males speaker respectively).

System 2 - iXtractor I-vector system was published in [23] and is closely related to the JFA framework. While JFA effectively splits model parameter space into wanted and unwanted variability subspaces, i-vector system aims at describing the subspace with the highest overall variability. If Eq. 5.1.3 characterizes JFA, then Eq. 6 characterizes the i-vector system:

$$\mathbf{M} = \mathbf{m} + \mathbf{Ti}, \tag{6}$$

where $\mathbf{T}$ is the subspace matrix, referred to as *i-vector extractor or* ixtractor

The ixtractor is trained using the same EM procedure as the subspace matrices in JFA with every segment being treated as a unique speaker. This way, i-vector system serves as a front-end or "feature extractor" for further processing, in which channel effects can be treated. In our case, we used LDA and Within-Class Covariance Normalization to transform the i-vectors to get rid of the unwanted variability.

When scoring a trial, such i-vector was estimated both for the enrollment part and the test segments. Scoring is therefore understood as comparing two i-vectors and the problem is symmetrical.

In our case, cosine distance of the i-vectors was taken as a score, i.e. the i-vectors were normalized to unit length and their dot product was taken as the score (see [23] for details).

5.1.4 Authentication

System 1 - Joint Factor Analysis - SVM We derived 300 speaker factors using JFA for each utterance and use them as a supervector to train SVM (Support Vector Machines). The background cohort for SVM are data from MOBIO database denoted as world-set. We used libsvm for all experiments with SVM [19].

System 2 - iXtractor We used gender independent UBM for this system. The iXtractor is trained on the same data as UBM. LDA and WCCN matrix is trained on the same data as UBM and MOBIO word-set data.

5.1.5 Normalization/Calibration

The score normalization was applied only to iXtractor system. We used s-norm normalization [23] with cohort derived from MOBIO word-set.

The experiments with SVM shows that the score normalization do not bring big improvement with this topology of the system.

Both systems are calibrated with Linear Logistic Regression (LLR) to produce true Log Likelihood Ratio score. Only a shift and scale is estimated to calibrate the scores. For convenience, FoCal toolkit by Niko Brummer[2] was used.

System 3 - Fusion We used Linear Logistic Regression (LLR) for training a linear fusion on development data of MOBIO database. At first the separate score were calibrated to produce Likelihood Ratio and then two shifts and one scale were trained. The fusion is linear and gender independent. We used this simple fusion, because we were afraid of over-training to the development data.

5.1.6 Discussion of Results

The results obtained for the two audio recognition systems are consistent across both the development and test sets. A summary of the HTERs can be found in Table 12. We see that there is an improvement with fusion about 10% relative against the better system. We decided to participate with the fusion of the two audio systems, because we saw consistent complementarity of the two systems. One was better for female and one for male so the fusion was ideal to preserve performances of both systems.

5.2. University of Avignon (LIA)

[2] http://niko.brummer.googlepages.com/focalbilinear

	Male	Female	Average
System 1	11.30%	12.37%	11.84%
System 2	12.55%	12.63%	12.59%
System 3	10.47%	10.85%	10.66%

Table 12. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

The LIA submitted two systems, systems 1 and 2, (respectively denoted LIA_A1 & LIA _A2) to the MOBIO contest. Both are based on the UBM/GMM (Universal Background Model / Gaussian Mixture Model) paradigm. During this evaluation, development, calibration and training (even for UBM training) were processed by only using the MOBIO corpus.

5.2.1 Feature Extraction

The two LIA systems use different LFCC parameterizations, both based on filter-bank analysis:

- LFCC48: the LIA_A1 system is based on 50 filter bank LFCC computed over 20ms Hamming windowed frames on the original 48kHz signal at a 10ms frame rate. Features are composed of 29 LFCC coefficients augmented with their 29 delta, 11 first double delta coefficients and the delta energy. Each acoustic vector is so composed of 70 coefficients.

- LFCC16: the LIA_A2 system is based on 24 filter bank LFCC computed over 20ms Hamming windowed frames on the 16kHz down-sampled signal at a 10ms frame rate. Features are composed of 19 LFCC coefficients augmented with the 19 delta, 11 first double delta coefficients and the delta energy. Each acoustic vector is so composed of 50 coefficients. Moreover, the bandwidth is limited to the 300-3400Hz range.

Finally, the acoustic vectors are normalised to fit a 0-mean and 1-variance distribution. The mean and variance estimators used for the normalisation are computed file by file on a set of frames selected using the process described in the next paragraph.

5.2.2 Voice Activity Detection and Speech Segmentation

The energy coefficients are first normalised using a mean removal and variance normalisation in order to fit a 0-mean and 1-variance distribution and then used to

train a three components GMM, which aims at selecting informative frames [6]. This approach aims to classify acoustic frames depending on the acoustic energy. Only frames corresponding tho the high-energy Gaussian components are labeled *speech*, others features are considered as not relevant.

After this first feature labelling, final morphological rules are applied on speech segments to avoid too short ones, adding or removing some speech frames applied in order to refine the speech segmentation.

5.2.3 Enrolment

For the two, previously described, parameterizations, UBM are trained using only the MOBIO UBM-set. Resulting world models are gender-dependent GMM with diagonal covariance matrices.

- the UBM of LIA_A1 system is a 512 GMM;
- the UBM of LIA_A2 system is a 256 GMM.

For a better separation of initial classes, frames are randomly selected among the entire learning signal via a probability followed by an iteration of the EM algorithm, to estimate the GMM parameters. During all the process, a variance flooring is applied so that no variance value is less than 0.5.

5.2.4 Authentication

The speaker models are adapted from the UBM via a MAP [53] adaptation. The relevant factor is fixed to 14. The score computation follows a classical log-likelihood computation using a $topN$ Gaussian computing.

5.2.5 Normalisation/Calibration

For both LIA GMM-UBM based systems, 211 male segments and 84 female segments from the MOBIO UBM-set are used as background data for a T-norm [4] score normalisation. Even if, the literature presents the ZT-norm as the reference normalisation, in the specific case of MOBIO better results were obtained by using only the T-normalisation, we assume that is probably due to the impostor cohort selected for score normalisation.

5.2.6 Discussion of Results

Results obtained with LIA_A1 and LIA_A2 systems on the test set are relatively better than the one obtained during the development phase. This can probably be explained by the similarity between the UBM-set and respectively the development and test sets. The GMM/UBM performance is strongly linked to the representativity of the UBM-set used for both UBM training and score normalisation. In this case, test-set seems closer from the UBM-set than the development set.

Finally, the state-of-the-art LIA speaker recognition system [37] is based on the Latent Factor Analysis (LFA) approach [38] which is known to be less performant than Joint Factor Analysis [30] approaches in case of short duration test segments. During the development phase, it seems that session's duration from the UBM-set and development set were too short to strongly estimate the LFA statistics.

	Male	Female	Average
System 1	14.74%	15.83%	15.29%
System 1 (norm)	14.49%	15.70%	15.10%
System 2	25.04%	18.59%	21.82%
System 2 (norm)	26.17%	19.77%	22.97%

Table 13. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

5.3. Tecnologico de Monterrey, Mexico and Arizona State University, USA (TEC-ASU)

The system we developed, named TECHila, evolved from our earlier systems that had used alternative data sets (YOHO, SV-TIMIT, and NIST2008) and it is based on the Gaussian Mixture Model (GMM) framework. TEChila aims to perform on par with state of the art methods for SRE such as [46], as well as to identify opportunities for improvements that may have been overlooked.

Most of the components were coded in Matlab for clarity and ease of inspection.

System 1 used ...

System 2 used ...

5.3.1 Voice Activity Detection and Speech Segmentation

The speech signal was downsampled to 8 Khz. Subsequently, a 25 ms analysis overlapping Hamming window, 10 ms frame rate, and pre-emphasis coefficient of .97 was applied. For a given conversation side, every frame log-energy was tagged as high, medium and low. Instead of a traditional voice activity detector, we used a frame removal technique. The low and 80% of the medium log-energy frames were then discarded. It is

important to note that the delta and double delta coefficients were obtained after the silent frames were removed. This 80% threshold is a heuristic that was derived empirically.

5.3.2 Feature Extraction

A short-time 256-pt Fourier analysis is performed on each overlapping window. The magnitude spectrum was transformed to a truncated vector of Mel-Frequency Cepstral Coefficients (MFCC), and a 23 channel filterbank. Following this step, we used two feature extraction approaches. In the first approach, the feature vector consisted of 33 attributes: 16 static Cepstral, 1log Energy, and 16 delta Ceptral coefficients. The second approach consisted of 49 attributes: 16 static Cepstral, 1log Energy, 16 delta Ceptral coefficients, and 16 double delta Ceptral coefficients.

Further, we implemented a feature warping algorithm on the obtained features. Feature warping belongs to the family of Gaussianization methods [45, 20] of normalization. The underlying idea in this normalization scheme is that every spectral attribute (Cepstral coefficient in our case) is normally distributed across time, and that the transmission channel distorts such a distribution. The task of feature warping is to undo the distortion caused by the channel by warping each attribute's scale so that the resulting attribute has a normal distribution. Traditionally, this warping is accomplished by first assembling an empirical CDF (cumulative distribution function) from the ranked features within 1.5 seconds before and after the current frame (3 seconds total), and then perform the CDF-inverse at the current frame.

5.3.3 Enrolment

A GMM (Gaussian mixture model) approach was adopted in this work. The evaluation was done independently for each gender, since it is reasonable to assume that each identity claim comes with a gender attribute. A gender-dependent and target-independent 512-mixture GMM anti-model model was trained from a pool of the MOBIO speech database. The EM (expectation maximization) algorithm was used to obtain the maximum likelihood estimates of the GMM parameters. TECHila's implementation of the EM algorithm for GMM uses the MPI (Message Passing Interface) environment to take full advantage of parallel computing infrastructure.

The GMM is first initialized using the K-means algorithm to obtain a set of 512 centroids. By using the k-means algorithm, the convergence of the EM is known to be faster. However, it is always important to check that the local bounds are not very restrictive, so that EM can make a satisfactory estimation. The EM is then repeated after the model had converged (about 3-5 iterations).

5.3.4 Authentication

A gender-dependent and target-independent 512-mixture GMM anti-model [32] model was trained from a pool of the MOBIO speech database (4893 audio files for male, 1764 for female). Target-dependent models were then obtained with a traditional MAP (maximum a posteriori) speaker adaptation [28]. Subsequently, two approaches were studied. In the first one, we used only one file from each speaker to train each target model (the average time of these utterances is 7 seconds). On the other one, we used the pool of all target files to compute each model.

The target-models are obtained with a traditional MAP (maximum a posteriori) speaker adaptation. The score obtained for every trial follows the hypothesis test framework, where the null hypothesis accepts the speaker as legitimate and the alternative hypothesis rejects him/her. Under this framework, the score is given the log likelihood ratio of two models: target-model and anti-model. As mentioned earlier, in the current implementation, the anti-model is target-independent.

5.3.5 Normalization/Calibration

No normalization of the scores was performed in this work.

5.3.6 Discussion of Results

The results obtained using our approach are summarized as follows:

- *Development database:* (33 attributes: 16 static Cepstral, 1log Energy, and 16 delta Ceptral coefficient, single file adaptation). Our best results (the first approach among the two mentioned in the authentication section) showed the following EER: 20.552% for male, 25.227% for female and a total average of 22.88%.

- *Test database:* (49 attributes: 16 static Cepstral, 1log Energy, 16 delta Ceptral coefficient, 16 double delta coefficient, all file adaptation). We obtained a EER of 15.453% for female, 17.414% for male and a total average of 16.45%.

We believe that our development results are higher because of the lack of the double delta coefficients, and the MAP training using a single file. We will consider

further normalization techniques in order to obtain better results as part of our future work.

	Male	Female	Average
System 1	20.55%	25.23%	22.89%
System 2	15.45%	17.41%	16.43%

Table 14. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

5.4. University of West Bohemia (UWB)

Our effort was to examine functionality of a system composed of several subsystems based on generative and discriminative models. We have utilized only the data provided by MOBIO.

System 1 used ...

System 2 used ...

System 3 used ...

System 4 used fusion ...

5.4.1 Voice Activity Detection and Speech Segmentation

In the pre-processing stage the speech signal was downsampled to 16 kHz and processed with a Voice Activity Detector (VAD) in order to discard non-speech frames. VAD was based on a set of filter-bank energy detectors situated in the frequency domain. Firstly, local Speech to Noise Ratios (SNRs) were computed for each frame as a mean of SNR estimated for each of the filter-banks. Second, global SNR was estimated (across whole utterance) as the mean value of local SNRs. At the end, frames with local SNRs higher than the global SNR were kept, all the other frames were discarded (marked as non-speech).

5.4.2 Feature Extraction

Our system exploited Mel Frequency Cepstral Coefficients (MFCCs) with 50 filter-banks. MFCCs were extracted each 10 ms utilizing a 25 ms hamming window, the C0 coefficient and energy were discarded, delta's were added, simple mean and variance normalization was applied and final set of features was downsampled with a factor 2. The final dimension of feature vectors reached 40.

5.4.3 Enrolment

Four systems were proposed, namely UWB_A1 : system based on Gaussian Mixture Models (GMMs) [52], system based on Support Vector Machines (SVMs) utilizing two types of kernels - UWB_A2 : GMM Supervector (GSV) kernel [14] and UWB_A3 : Generalized Linear Discriminant Sequence (GLDS) kernel [13], and finally, UWB_F : system based on fusion of mentioned approaches. GMMs were adapted from a Universal Background Model (UBM) with 510 mixtures trained on all the gender specific data provided by MOBIO and denoted as world-set, hence genders were handled separately. Maximum A-Posteriory (MAP) adaptation was performed with a relevance factor 14, and only means were adapted. UBM was trained using Maximum Likelihood (ML) estimation, which was preceded by Distance Based (DB) algorithm in order to initialize the ML training. The GSV kernel made use of concatenated GMM means, hence a 20400 dimensional supervector (SV) was formed. Polynomial order 3 was assumed by construction of GLDS supervectors resulting in SV dimension of 12341. Impostors for SVM modeling were also drawn from the world-set in a gender specific manner.

5.4.4 Authentication

In the case of GMM system the Log-Likelihood Ratio (LLR) approach was used to score the trials, and in the case of SVM models a simple scalar multiplication was utilized. In order to fuse the results of individual systems a linear weighing of particular scores was performed. Weights were trained on the development set according to a simple gradient method with auxiliary function given as overall Equal Error Rate (EER) of fused results.

5.4.5 Normalization/Calibration

UWB systems did not use score normalization as no data were found to be suitable for such a task. Some efforts were made to enroll the world-set, but the results obtained on the development set were unconvincing. However, it turns out that SVM systems perform well regardless the TNorm [15], which is in the case of SVM of minor importance.

5.4.6 Discussion of Results

Results obtained on the development and test set are similar. Decrease of the performance was observed for the UWB_A2 system, mainly for female tests. It is well known that SVM training demands a lot of background

data to be trained, especially in cases of one-versus-all training utilizing high dimensional SVs. Our system used as impostors speakers from the world-set provided by MOBIO, where only 14 female speakers and 39 male speakers were present. Each of the speakers was represented with multiple session recordings processed separately and used as an impostor regardless of the pertinence to the same speaker (in common, 1764 female impostors and 4893 male impostors were used). Still, one can not assume that a discriminative system trained just on a few speakers could generalize well to unseen data, anyhow it can bring some additional information utilized in advance in score fusion. Best performance was achieved with the GMM system UWB_A1, hence a conclusion can be made that a UBM-GMM system is the best answer in situations where only few data for training are available.

	Male	Female	Average
System 1	9.76%	10.73%	10.24%
System 2	19.08%	14.46%	16.77%
System 3	12.03%	11.33%	11.68%
System 4	11.18%	10.00%	10.59%

Table 15. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

5.5. Swansea University and Validsoft (SUV)

The speaker verification systems submitted by Swansea University and Validsoft are based on standard Gaussian Mixture Models (GMMs) [51], whose originality lies in the use of wide band feature extractors, an idea already explored by Swansea University during the Biosecure evaluation campaign [26]. They were developed using SPro [3] and ALIZE [9] open source toolkits. The GMM systems are as described in [8] and the front-end is an adaptation from the mean-based feature extraction described in [27].

System 1 used ...

System 2 used ...

System 3 used fusion ...

[3] http://gforge.inria.fr/projects/spro/

5.5.1 Voice Activity Detection and Speech Segmentation

Voice activity detection is a simple approach based on energy distributions. The threshold is set on the mean of the Gaussian of highest energy out of three Gaussians fitted with EM on the energy components.

5.5.2 Feature Extraction

Two types of frontends were used, $F1$ and $F2$, both cepstral coefficient based. No downsampling was performed. On such a wide band Mel-frequency cepstral coefficients (MFCCs) were found to perform better than linear ones (LFCCs). The difference between the two frontends come from the number of filter bands and the number of coefficients kept after discrete Cosine transform (DCT). $F1$ and $F2$ configurations are as follow:

- $F1$: MFCC, 50 bands, 29 DCT coefficients, 29 delta + delta Energy
- $F2$: MFCC, 24 bands, 16 DCT coefficients, 16 delta + delta Energy

Apart from the fact that the filters are spread on a wide band (0 Hz-24 kHz) $F2$ corresponds to a standard MFCC configuration. With a larger number of filter bands $F1$ was found to performed better and is used in the system SUV_A1. SUV_A2 is a fusion of two systems whose only difference are the frontends, one based on $F1$ and the other one on $F2$.

After speech activity detection, 0-mean 1-variance normalisation is performed on the full utterance.

5.5.3 Enrolment

Enrolment is based on conventional GMM with MAP adaptation of the Gaussian components. The Universal Background Models (UBMs) used are gender dependant and trained with all the data from MOBIO "devset". GMMs have 512 components. The relevance factor is 14 and no channel normalisation is used.

5.5.4 Authentication

Speaker authentication tests are standard log-likelihood ratio computations.

5.5.5 Normalization/Calibration

T-normalisation is performed with gender dependant cohorts chosen randomly in MOBIO 'world-set' (158 for female, 182 for male).

For SUV_A2, score-level fusion of the two GMM systems is done with equal weights after T-norm.

5.5.6 Discussion of Results

SUV submission is based on a standard GMM-UBM approach. Due to the limited size of the development set, no attempt was made to use more sophisticated approaches such as SVM or Factor Analysis. Results on the test set are in line with the results on the development set with actually some improvement on the female subset. By recording speaker directly on the handset MOBIO provided a database of speech sampled at 48 kHz in contrast with the majority of other speech databases usual recorded on 8 kHz telephony speech. This was the main motivation and the main focus of development for SUV submission. Results on the development set suggests that working on wider band bring relative improvements of performance in the region of 20 to 30 %. Further work is needed to contrast and compare systems based on telephony speech which now use recordings from $1000s$ of speakers and other "high-quality speech" based systems whose background data is limited to a few 10s of speakers.

	Male	Female	Average
System 1	14.70%	16.00%	15.09%
System 1 (norm)	14.04%	15.42%	14.73%
System 2	15.09%	17.81%	16.45%
System 3	13.57%	15.27%	14.42%

Table 16. Table presenting the final results (HTER) on the Test set for the MOBIO Phasel database.

6. Discussion

A quick summary of the uni-modal face systems can be found in Table 17. The results are also presented in the DET plots in Figure 1 (male trials) and 2 (female trials). The majority of the systems use an OpenCV like face detection scheme and all seem to have similar performance. The systems which use an alternative face detection (Visidon, UNIS and ITI) scheme seem to have a definite advantage over those who don't.

The impact of the face detection algorithm can be seen clearly when examining the two systems from ITI. The difference between the two systems from ITI come only from the use of a different face detection technique: System 1 uses the frontal OpenCV face detector and System 2 uses the OmniPerception SDK. The difference in face detector alone leads to an absolute improvement of the average HTER of more than 4%. This leads us to conclude the one of the biggest challenges for video based face recognition is the problem of accurate face detection.

	Male	Female	Average
IDIAP*	25.45%	24.39%	24.92%
ITI*	16.92%	17.85%	17.38%
NICTA*	25.43%	20.83%	23.13%
TEC*	31.36%	29.08%	30.22%
UNIS*	9.75%	12.07%	10.91%
VISIDON*	10.30%	14.95%	12.62%
UON*	29.80%	23.89%	26.85%
NTU*	20.50%	27.26%	23.88%
UPV*	21.86%	23.84%	22.85%

Table 17. Table presenting the final results for face recognition on the Test set for the MOBIO Phasel database.

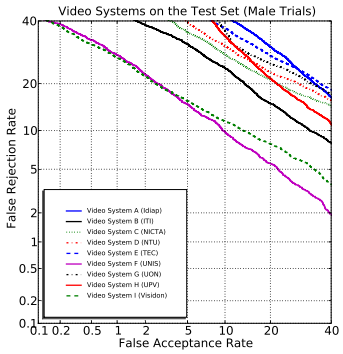

Figure 1. Male DET for the Video.

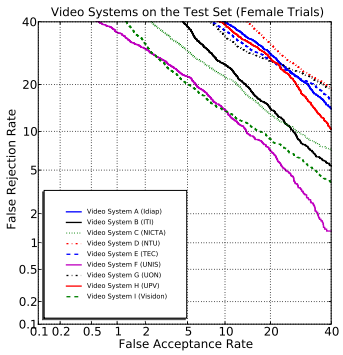

Figure 2. Female DET for the Video.

A quick summary of the HTERs for the uni-modal speaker systems can be found below in Table 18.

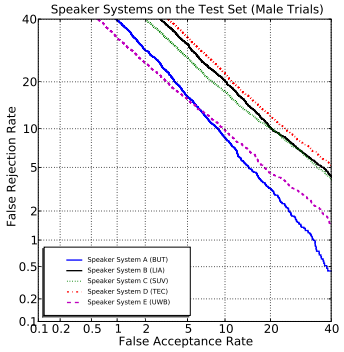

Figure 3. Male DET for the Audio.

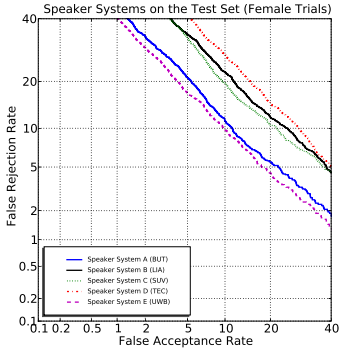

Figure 4. Female DET for the Audio.

	Male	Female	Average
BUT*	10.47%	10.85%	10.66%
LIA*	14.49%	15.70%	15.10%
SUV*	13.57%	15.27%	14.42%
TEC*	15.45%	17.41%	16.43%
UWB*	11.18%	10.00%	10.59%

Table 18. Table presenting the final results for speaker recognition on the Test set for the MOBIO Phasel database.

To examine the possibility of fusing the two modalities we took two of the better systems from each modality and tried to fuse them. We took pairs of systems from each modality and fused them, this led to four possible fusion system which are listed in Table 19. Fusing the best combination (Face1 + Speaker1) we obtained the a HTER of 3.00% and 5.50% on the Test set (which is significantly better than either system on its own), the results of this fusion are summarised by a DET plot 5 and two EPCs **??**.

	Fusion	
	Male	Female
Face1 + Speaker1	2.22%	2.13%
Face1 + Speaker2	3.80%	2.80%
Face2 + Speaker2	1.78%	4.13%
Face2 + Speaker1	3.11%	4.67%

Table 19. Table presenting the initial fusion results on the Development set for the MOBIO Phasel database in term of equal error rate.

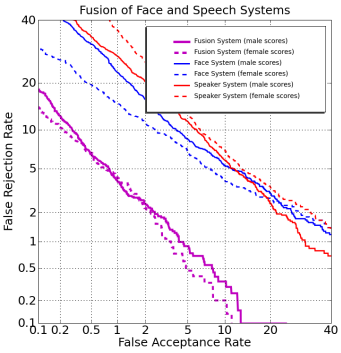

Figure 5. DET Plot for the fusion on the Test set.

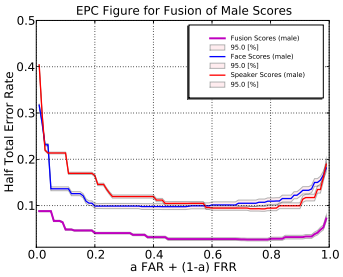

Figure 6. EPC for male scores on the fusion data on the Test set.

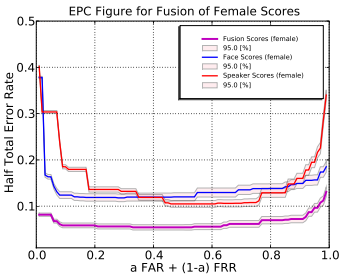

Figure 7. EPC for female scores on the fusion data on the Test set.

7. Conclusion

8. Acknowledgements

This work has been performed by the MOBIO project 7th Framework Research Programme of the European Union (EU), grant agreement number: 214324. The authors would like to thank the EU for the financial support and the partners within the consortium for a fruitful collaboration. For more information about the MOBIO consortium please visit `http://www.mobioproject.org`. The authors would also like to thank Phil Tresadern (University of Manchester), Bastien Crettol (Idiap Research Institute), Norman Poh (University of Surrey), Christophe Levy (University of Avignion), Driss Matrouf (University of Avignon), Timo Ahonen (University of Oulu), Honza Cernocky (Brno University of Technology) and Kamil Chalupnicek (Brno University of Technology) for their work in capturing this database and development of the protocol.

References

[1] Visidon ltd. (http://www.visidon.fi).

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face Recognition with Local Binary Patterns. *LECTURE NOTES IN COMPUTER SCIENCE*, pages 469–481, 2004.

[3] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol. Face recognition using hog-ebgm. *Pattern Recognition Letters*, 29(10):1537–1543, 2008.

[4] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score Normalization for Text-Independent Speaker Verification System. *Digital Signal Processing*, 1(10):42–54, 2000.

[5] I. M. Author. Some related article I wrote. *Some Fine Journal*, 99(7):1–100, January 1999.

[6] L. Besacier, J.-F. Bonastre, and C. Fredouille. Localization and selection of speaker-specific information with statistical modeling. *Speech Communication*, 31(2-3):89–106, 2000.

[7] W. W. Bledsoe. The model method in facial recognition. Technical report, Panoramic Research Inc., 1966.

[8] J. Bonastre, N. Scheffer, C. Fredouille, and D. Matrouf. NIST04 speaker recognition evaluation campaign: new LIA speaker detection platform based on ALIZE toolkit. In *Proceedings of NIST speaker recognition workshop*, 2004.

[9] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason. ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In *Proceedings Odyssey - The Speaker and Language Recognition Workshop*, 2008.

[10] G. R. Bradski. Computer video face tracking for use in a perceptual user interface. *Intel Technology Journal*, Q2, 1998.

[11] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky. Analysis of feature extraction and channel compensation in GMM speaker recognition system. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):1979–1986, Sept. 2007.

[12] J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9), Sept. 1997.

[13] W. Campbell. Generalized linear discriminant sequence kernels for speaker recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP'02*, 1:I–161–I–164, 2002.

[14] W. Campbell, D. Sturim, and D. Reynolds. Support vector machines using gmm supervectors for speaker verification. *Signal Processing Letters, IEEE*, 13(5):308–311, 2006.

[15] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, 1:I–I, 2006.

[16] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of mlp and gmm classifiers for face verification on xm2vts. In *International Conference on Audio- and Video-based Biometric Person Authentication*, pages 1058–1059, 2003.

[17] C. Chan, J. Kittler, N. Poh, T. Ahonen, and M. Pietikäinen. (multiscale) local phase quantization histogram discriminant analysis with score normalisation for robust face recognition. In *VOEC*, pages 633–640, 2009.

[18] C.-H. Chan, J. Kittler, and K. Messer. Multi-scale local binary pattern histograms for face recognition. In S.-W. Lee and S. Z. Li, editors, *ICB*, volume 4642 of *Lecture Notes in Computer Science*, pages 809–818. Springer, 2007.

[19] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.

[20] Chen and R. Gopinath. Gaussianization. *NIPS*, 2000.

[21] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR2000)*, pages 142–149, 2000.

[22] A. Das. Audio visual person authentication by multiple nearest neighbor classifiers. In *SpringerLink*, 2007.

[23] N. Dehak, R. Dehak, P. Kenny, N. Brmmer, P. Ouellet, and P. Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Proc. International Conferences on Spoken Language Processing (ICSLP)*, pages 1559–1562, Sept. 2009.

[24] H. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen. Multi-modal person identification in a smart environment. In *IEEE CVPR*, 2007.

[25] A. N. Expert. *A Book He Wrote*. His Publisher, Erewhon, NC, 1999.

[26] B. Fauve, H. Bredin, W. Karam, F. Verdet, A. Mayoue, G. Chollet, J. Hennebert, R. Lewis, J. Mason, C. Mokbel, and D. Petrovska. Some results from the biosecure talking face evaluation campaign. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2008.

[27] B. Fauve, N. W. D. Evans, and J. Mason. Improving the performance of text-independent short duration GMM- and SVM-based speaker verification. In *Proceedings Odyssey - The Speaker and Language Recognition Workshop*, 2008.

[28] J. Gauvain and C. Lee. Map estimation of continuous density hmm: Theory and applications. *DARPA Sp. & Nat. Lang. Workshop*, February 1992.

[29] N. Hieu, L. Bai, and L. Shen. Local gabor binary pattern whitened pca: A novel approach for face recognition from single image per person. In *The 3rd IAPR/IEEE International Conference on Biometrics, 2009. Proceedings.*, 2009.

[30] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio Speech and Language Processing*, 15(4):1435, 2007.

[31] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of inter-speaker variability in speaker verification. In *IEEE Transactions on Audio, Speech and Language Processing*, July 2008.

[32] C.-H. Lee. A unified statistical hypothesis testing approach to speaker verification and verbal information verification. *invited paper in Proc. COST Workshop on Speech Technology in the Public Telephone Network: Where are we today?*, September 1997.

[33] P.-H. Lee, G.-S. Hsu, and Y.-P. Hung. Face verification and identification using facial trait code. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1613–1620, 2009.

[34] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *DAGM'03, 25th Pattern Recognition Symposium*, pages 297–304, Madgeburg, Germany, 2003.

[35] C. Liu. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:725–737, 2006.

[36] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[37] D. Matrouf, J.-F. Bonastre, C. Fredouille, A. Larcher, S. Mezaache, M. McLaren, and F. Huenupan. LIA GMM-SVM system description: NIST SRE08. In *NIST Speaker Recognition Evaluation Workshop*, Montreal (Canada), april 2008.

[38] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *Proc.*

Interspeech 2007, International Conference on Speech Communication and Technology, 2007.

[39] P. Matějka, L. Burget, P. Schwarz, and J. Černocký. Brno University of Technology System for NIST 2005 Language Recognition Evaluation. In *IEEE Odyssey: The Speaker and Language Recognition Workshop*, pages 57–64, San Juan, Puerto Rico, June 2006.

[40] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostin, F. Cardinaux, S. Marcel, S. Bengio, C. Sanderson, N. Poh, Y. Rodriguez, J. Czyz, L. Vandendorpe, C. McCool, S. Lowther, S. Sridharan, V. Chandran, R. P. Palacios, E. Vidal, L. Bai, L. Shen, Y. Wang, C. Yueh-Hsuan, L. Hsien-Chang, H. Yi-Ping, A. Heinrichs, M. Muller, A. Tewes, C. von der Malsburg, R. Wurtz, Z. Wang, F. Xue, Y. Ma, Q. Yang, C. Fang, X. Ding, S. Lucey, R. Goss, and H. Schneiderman. Face authentication test on the banca database. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 4, pages 523–532, 2004.

[41] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, and N. Mavity. Face verification competition on the xm2vts database. In *AVBPA*, pages 964–974, 2003.

[42] D. Monzo, A. Albiol, and J. Sastre. Hog-ebgm vs. gabor-ebgm. In *International Conference on Image Processing*, pages 1636–1639, October 2008.

[43] D. Monzo, A. Albiol, J. Sastre, and A. Albiol. Precise eye localization using hog descriptors. *Under review*.

[44] Neurotechnologija. Verilook SDK. Neurotechnologija Biometrical and Artificial Intelligence Technologies (http://www.neurotechnologija.com).

[45] J. Pelcanos and S. Sridharan. Feature warping for robust speaker verification. *2001: A Speaker Odyssey Workshop*, June 2001.

[46] A. E.-H. Petrovska-Delacretaz and G. Chollet. Text-independent speaker verification: State of the art and challenges. *LNCS Springer*, May 2007.

[47] J. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Conference of Computer Vision and Pattern Recognition*, volume 1, pages 947–954, 2005.

[48] J. P. Phillips, H. Moon, S. Rizv, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[49] N. Poh, C. H. Chan, J. Kittler, S. Marcel, C. McCool, E. A. Rúa, J. L. A. Castro, M. Villegas, R. Paredes, V. Struc, N. Pavesic, A. A. Salah, H. Fang, and N. Costen. Face video competition. In *3rd International Conference on Biometrics (ICB)*, volume 5558 of *LNCS*, pages 715–724. Springer, Alghero, (Italy), June 2009.

[50] D. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 963–966, 1997.

[51] D. A. Reynolds, T. Quatieri, and R. Dunn. Speaker recognition using adapted mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[52] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19 – 41, 2000.

[53] D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 3(1):72–83, January 1995.

[54] Y. Rodriguez. *Face Detection and Verification using Local Binary Patterns*. PhD thesis, EPFL, 2006.

[55] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[56] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International Conference on Biometrics, Lecture Notes in Computer Science (LNCS)*, volume 5558, pages 199–208, 2009.

[57] P. Schwarz, P. Matějka, and J. Černocký. Hierarchical structures of neural networks for phoneme recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 325–328, Toulouse, France, May 2006.

[58] S. Shan, W. Zhang, Y. Su, X. Chen, W. Gao, I. FR-JDL, and B. CAS. Ensemble of Piecewise FDA Based on Spatial Histograms of Local (Gabor) Binary Patterns for Face Recognition. In *Proceedings of the 18th international conference on pattern recognition*, pages 606–609, 2006.

[59] E. Shriberg, L. Ferrer, and S. Kajarekar. Svm modeling of snerf-grams for speaker recognition. In *International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, Oct. 2004.

[60] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman. MLLR transforms as features in speaker recognition. In *International Conference on Spoken Language Processing (ICSLP)*, pages 2425–2428, Lisbon, Portugal, Sept. 2005.

[61] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *AMFG 2007*, volume 4778, pages 168–182, 2007.

[62] X. Tan and B. Triggs. Fusing gabor and lbp feature sets for kernel-based face recognition. In *AMFG*, 2007.

[63] M. Villegas and R. Paredes. Illumination invariance for local feature face recognition. In *1st Spanish Workshop on Biometrics*, Girona (Spain), June 2007.

[64] M. Villegas and R. Paredes. Simultaneous learning of a discriminative projection and prototypes for nearest-neighbor classification. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[65] M. Villegas, R. Paredes, A. Juan, and E. Vidal. Face verification on color images using local features. *Computer Vision and Pattern Recognition Workshops, 2008. CVPR Workshops 2008. IEEE Computer Society Conference on*, pages 1–6, June 2008.

[66] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features, 2001.

[67] P. A. Viola and M. J. Jones. Robust real-time face detection. volume 57, pages 137–154, 2004.

[68] L. Wiskott, J. M. Fellous, N. Kruger, and C. Malsburg. Face recognition by ebgm. Technical report, Ruhr-Universitat Bochum, April 1996.

[69] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Real-Life Images workshop at the European Conference on Computer Vision (ECCV)*, October 2008.

[70] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition. In *Proc. ICCV*, pages 786–791, 2005.