Multi-modal person verification tools using speech and images[*]

M. Acheroy RMA-B

C. Beumier RMA-B

J. Bigün[†] EPFL-CH

G. Chollet IDIAP-CH

B. Duc EPFL-CH

S. Fischer EPFL-CH

D. Genoud IDIAP-CH

P. Lockwood MATRA-F

G. Maitre IDIAP-CH

S. Pigeon UCL-B

I. Pitas UAT-GR

K. Sobottka UAT-GR

L. Vandendorpe UCL-B

Abstract

We propose multi-modal person verification using voice and images as a solution to the secured access problem. The necessary i/o devices are now standard, cheaply available and, most importantly, constitute the two most important human communication modalities. The visual part currently involves i) matching of a coarse grid containing Gabor phase information from face images, ii) facial feature localization and extraction iii) 3D biometrical feature extraction by structured light. The acoustic part uses three methods (DTW,SOSM and HMM) to compare voice references extracted from the speech signal. In the acoustic part LPC coefficients are extracted and three different classifiers are used in parallel. The global decision is taken by applying a Furui threshold to the individual methods and in combining the individual results according to a majority law.

1 Introduction

In tele-services and tele-shopping applications, the usage scenarios are such that a large number of potential services are hampered from being put in operation.

[*] This work has been carried out within the framework of the European ACTS-M2VTS project.

[†] Author for correspondance: EPFL; DE-LTS; CH-1015; Lausanne

Many of those which are in operation or in planning rely heavily on the transfers of PIN's over the unprotected telephone lines. The credit card (which are around since several decades) theft is a nuisance for individuals as well as for the market which accepts them. Also with the current PIN based technologies, an application in which a specific user is given unlimited use of a service, e.g. consulting an updated document or database, is difficult to open to wide public as this would very quickly lend itself to abuse in that some users would voluntarily give away their pin. Tele banking services based on voice and video messaging could easily be implemented if the problem of secure verification was an integral part of the messaging system.

Various mono modal person verification methods are known and are partly already available on the market as products. They suffer however, from various drawbacks including high false acceptance rates, are perceived invasive, or simply too demanding in terms of resources. Cheap Mono-modal recognition techniques are likely to reach in a close future a saturation in performance. A promising, but challenging way of overcoming such limitations, consists in combining results from several modalities. The work presented in this paper is concerned with person identification/verification based on face and voice features extracted from visual and acoustic data. The choice of these modalities, [6], is motivated by the fact that these are part of the natural human messaging modalities, and are cheap, to the extent that they are offered as standard accessories of personal computers and work stations.

2 Database for tests

The goal of a multi-modal recognition scheme is to improve the recognition rates by combining single modalities, in this context face and voice features. This requires the development of fusion methods, i.e. the merge of individual results given by separate analysis of different modalities (e.g. voice texture and face features) or, more efficiently, the direct analysis of a combination of different modalities (e.g. study of speech/lip synchronization). Due to the relative novelty of multi-modal identification, our own material had to be recorded since no existing database could meet our requirements of offering all modalities needed by the multiple recognition tasks. These requirements are the presence of synchronized speech and image in the database and the possibility of extracting 3-D face features from the same database.

Our current database includes 37 different faces and provides 5 shots for each person. These shots were taken at one week intervals or when drastic face changes occurred in the meantime. During each shot, people are asked to count from '0' to '9' in their native language (most of the people is French speaking), rotate the head from 0 to -90 degrees, again to 0, then to +90 and back to 0 degrees. They are also asked to rotate the head again without glasses if they wear any. From the whole sequence, 3 parts are extracted : the "voice" sequence, the "motion" sequence and the "glasses off" motion sequence (if any). The first data sequence can be used for speech verification, 2-D dynamic face verification (choosing the most appropriate picture out of the sequence) and speech/lips movement correlation. The other two sequences are meant for face recognition purposes only and provide information about the 3-D face features thanks to the motion. These two sequences may also be used for implementing and comparing other recognition techniques like recognition from 2-D facial pictures, profile view and multiple views. For each person of the database, the most difficult shot to recognize is labeled as the 5th shot. They mainly differ from the others because of face variations (head tilted, eyes closed, different hairstyle, presence of a hat/scarf...), voice variations or shot imperfections (poor focus, different zoom factor, poor voice SNR...).

It was decided to use good quality material for the recording, leaving space in the future to degrade quality in order to simulate a given low-cost acquisition system. A Hi8 video camera (576x720, 50Hz-interlaced, 4:2:2) was chosen for the shooting and a D1 digital recorder for the recording and editing. In order to reduce the storage requirement, television sequences are then down-converted into CIF (288x360 pixels, 25Hz-Progressive, 4:2:2). This conversion removes every other field and performs horizontal down-sampling in the remaining frame with respect to the MPEG-2 TM5 specification. By keeping active pixels only, the final resolution for the database images is 286x350 pixels. Concerning voice acquisition, the sound track is digitally recorded using a 48kHz sampling frequency and 16 bit linear data.

Besides the particular case of the last shots, the database can be considered as having been produced under "ideal" shooting conditions (good picture quality, indoor shooting, nearly constant illumination, uniform gray background) and within a highly cooperative scenario (as much as they could, people followed the instructions they were given). Nevertheless, we can notice some impairments with respect to the theoretical case : - some people do no rotate their head properly (horizontal translation of the head in the direction of the rotation, vertical tilt depending on the rotation angle, no full covering of the 180 frontal degrees...), - some people might have their mouth open during one rotation of the head, closed during the other, ending up on different shapes in the profile view, - some people close their eyes while moving the head, - the direction of starting the rotation of the head is not, fixed over the different shots, - some people are speaking very low (resulting in a poor sound SNR), - some people can not keep from smiling during the shot, - rotation speed can be highly variable between different shots, but also within the same shots, - reflections on eyes and glasses, - blurry images during fast head rotation, due to limited shutter speed.

However, similar imperfections (combined with other as well) will appear when implementing a practical recognition scheme. Moreover people will expect the recognition algorithms to be able to deal with such imperfections. From this point of view, this face database can be seen as a good material to test the robustness of the recognition algorithms with regards to common problems. Assuming an algorithm would not overcome the imperfections encountered here, it would be difficult for this algorithm to overcome those associated with true operational conditions.

3 Verification by images

3.1 Extraction of facial features

In this approach the biometric features based on the distances between eyes and mouth, as well as the shape of the head constitute the recognition base. Due to variations in illumination, background, visual angle and facial expressions, the problem is complex. In the following we present an approach that localizes faces in color images on the base of shape and color (HSV) information. The hypotheses for faces are verified by searching for facial features inside of the face-like regions. This is done by applying morphological operations and minima localization to intensity images.

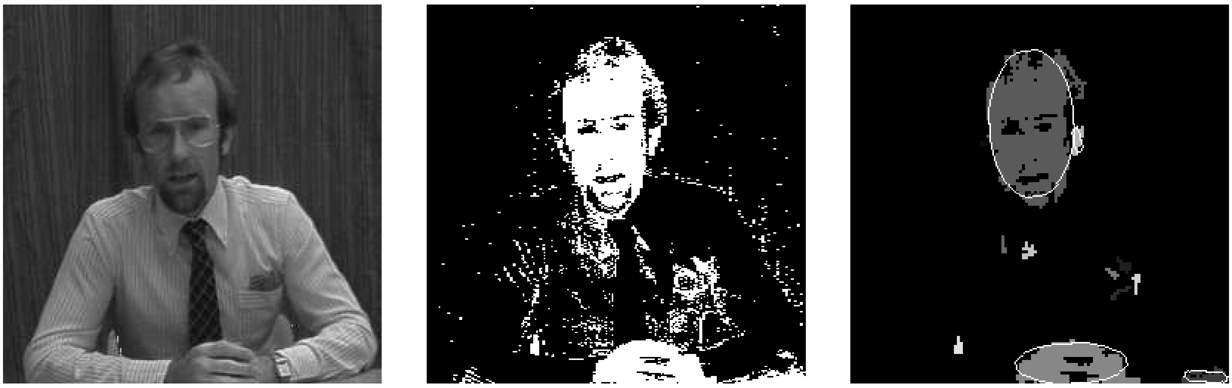
3.1.1 Face localization and approximation

In the field of face localization, approaches have been published using texture [7], depth [12], shape [11] and color information [24] or combinations of them. Still the detection of facial regions out of scenes with complex background is a problem.

In our approach we take advantage of the skin-specific color and the oval shape of faces. As discriminating color information we consider the attributes hue and saturation.

The effectiveness of using color information was also shown in [7]. Our results of the segmentation step are shown in Fig. 1a,b.

Because faces are characterized by their oval shape, looking for faces in an image means to detect objects with nearly elliptical shape. Mostly this is done based on edges (e.g. [11]), but we have chosen an approach based on regions. The advantage of considering regions is that they are more robust against noise and changes in illumination. In a first step connected components are determined by applying a region growing algorithm at a coarse resolution of the segmented image. Then we check for each connected component, if its shape is nearly elliptical or not. For that, based on moments, we compute the best-fit ellipse E and then we assess, how well the connected component C is approximated by E.



(a) (b) (c)

Figure 1: Detection of facial regions

This is done by evaluating the following measure V:

$$V = \frac{\sum_{(x,y)\in E}(1 - b(x,y)) + \sum_{(x,y)\in C\setminus E} b(x,y)}{\sum_{(x,y)\in E} 1} \tag{1}$$

with

$$b(x,y) = \begin{cases} 1 & \text{if } (x,y) \in C \\ 0 & \text{otherwise} \end{cases}$$

V determines the distance between the connected component and the best-fit ellipse by counting the "holes" inside of the ellipse and the points of the connected component that are outside of the ellipse. Based on a threshold on this ratio, the ellipses that are good approximations of connected components are selected and considered as face candidates. In the case of the example we obtain the results shown in Fig. 1c.

By searching for facial features inside of the connected components, the face hypotheses are verified.

3.1.2 Eyes and mouth localization

In intensity images eyes and mouth differ from the rest of the face because of the color of the pupils, the sunken eye-socket and the light red color of the lips. We make use of this observation and extract the positions of eyes and mouth by analysis of the topographic grey level relief inside of face-like regions.

In a preprocessing step, we enhance the dark regions by applying a grey scale erosion [19] and an extremum sharpening operation [18]. For the example scene, the results illustrated in Figure 2 (left) are obtained. Eyes and mouth and parts of the hair and beard regions are emphasized.

The position of eyes and mouth are determined by searching for minima in the topographic grey-level relief. For that we first compute the mean grey-level of

every row of the connected component and then determine minima in this y-relief. Beginning with the uppermost minima in y-direction, we search for minima in x-direction. The positions of the eyes are found, if two minima in x-direction are detected that meet the requirements for eyes concerning e.g. eye distance, relative position inside of head, significance of maximum between them. In case that reliably candidates for eyes are found, we look for mouth candidates. The search is started below of the eyes. Again, first the y-relief is investigated for minima and in case of a minimum, minima in x-direction are detected. For each of these minima is checked, if it meets the requirements for the mouth that concern e.g. width of the mouth, relative position inside of the head and relative position between eyes and mouth. On this basis, the best mouth candidate is chosen.
Examples for such reliefs in x- and y- direction are shown in Fig. 2.

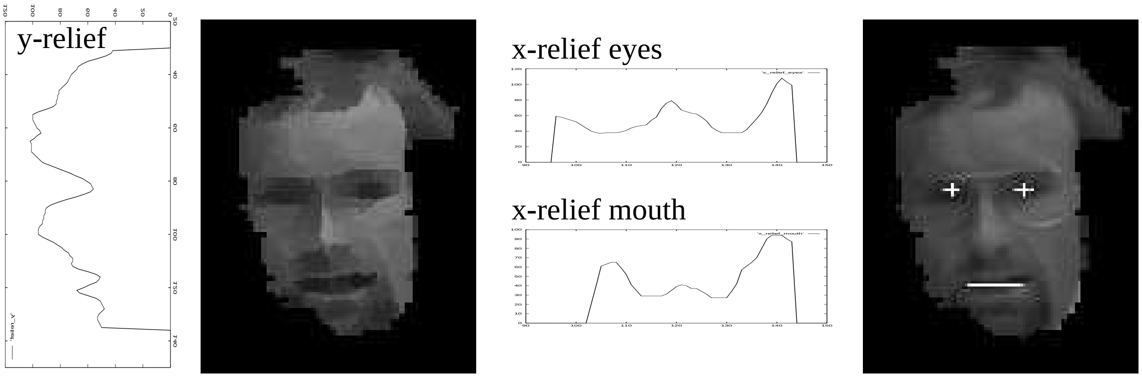


Figure 2: Detection of facial features

In the y-relief minima can be seen for eyes, mouth and beard. The minima in x-direction correspond with the vertical position of eyes and mouth.
Results of the detection of facial features are shown in the right. The eyes and the mouth are well localized.

3.2 Gabor response matching on a grid for face recognition

3.2.1 Feature vectors

The feature vectors we use are sets of features that describe local properties of points in the image. A feature vector can describe complex structures like line stops, corners, and crossings without having a model or any a priori knowledge of the structure being described.

Each face is described by a set of feature vectors positioned on nodes of a coarse grid, similar to that in [17]. Comparing two face images is accomplished by matching and adapting a grid taken from one image to the features of the other image.

Fig. 3 shows a feature vector that is used for every grid node as feature vector. We use complex Gabor responses to determine the feature vector from filters with 6 orientations and 3 resolutions.

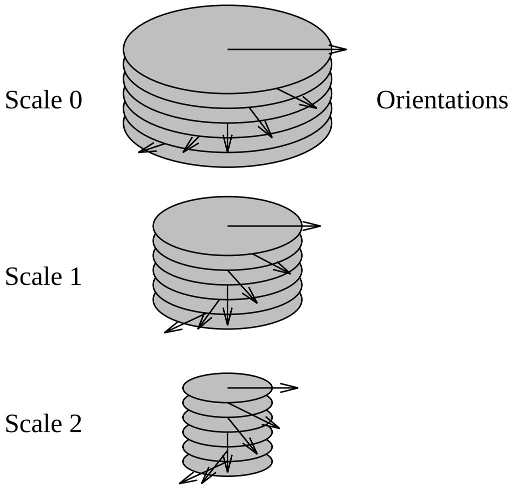


Figure 3: The feature vectors we have used are based on Gabor features from filters with 6 orientations and 3 resolutions.

We have examined two different feature vectors: phase and modulus of the complex valued vectors we obtained from Gabor responses.

3.2.2 Matching Gabor responses on a grid

The method for grid matching we employ consists of two steps. The first, coarse matching translates the undeformed grid that has been stored in a database and attempts to find the best correspondence between the grid and the image to be verified. The second step performs local matches around the grid points in order to find local minima. This results in a deformed grid. The degree of deformation as well as the mismatch of the feature vectors can be used as a measure for face difference. We have examined two different distance measures, namely the sum of all feature vector distances between grid and test image, and the sum of changes in Euclidean distances between the grid nodes caused by the deformation.

The first experiments suggest that the sum of the feature vector distances provides a reliable measure for pattern distance.

Other methods for calculating the grid distance could be based on the grey levels of the deformed rectangles that are being formed by the edges of the grid.

In this case the graph matching is only used to tackle the correspondence problem between two face images.

3.2.3 Performance Measures

A number of measures can be used to evaluate the performance of a recognition scheme. The performance measures that are used for person identification is the recognition rate (RR), false rejection rate (FRR), and false acceptance rate (FAR). The multi-modal approach requires a measure to evaluate the importance of every modality. A useful measure could be the ratio between the cost of an additional modality and the improvement in the recognition rate.

The performance should be measured on a standard face image database to allow for the comparison with other identification schemes. However the application areas of person identification are different, and the requirements vary considerably. In addition to that we need a database containing different modalities to evaluate the methods we have developed in a multi-modal person identification scheme. Hence, the need of a special face and speech database mentioned previously.

3.2.4 Results on grid matching

For face recognition, we have used an image database with 40 persons. Fig. 4 and Fig. 5 show two faces to be compared. Fig. 6 shows the same image as in Fig. 4 but with a superimposed test grid. Fig. 7 shows the deformed test grid computed from Fig. 6 that best matches the image in Fig. 5.

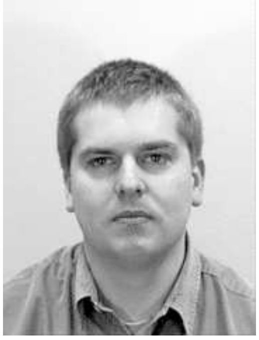


Figure 4: First test image.

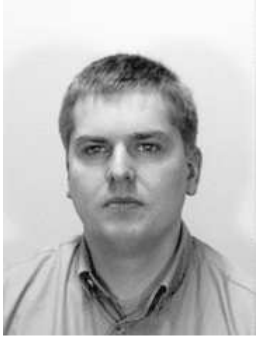


Figure 5: Second test image.

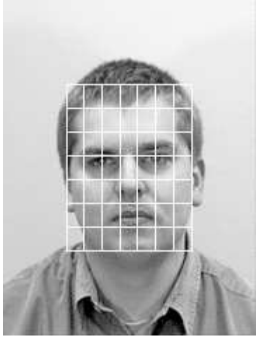


Figure 6: First test image with superimposed test grid.

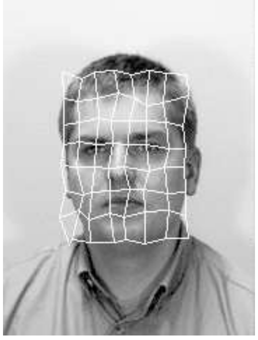


Figure 7: Second test image with deformed test grid.

Table 1 shows the distance measure between face images as a matrix. We have used pairs of images from 10 individuals not displaying any emotion, that have been taken at times more than two weeks apart. We can show that corresponding face images are recognised in most of the cases as more similar than images from different individuals. Indeed, one can notice that for every line and every column, the diagonal element is the smallest. This means that by taking the minimum distance as criterion for recognition, no error is made.

person no	1	2	3	4	5	6	7	8	9	10
1	1020	1410	1635	1430	1360	1513	1570	1516	1526	1530
2	1587	1034	1574	1587	1315	1309	1417	1459	1377	1403
3	1563	1550	1267	1400	1518	1537	1542	1445	1491	1652
4	1567	1428	1392	1274	1448	1448	1536	1438	1393	1584
5	1303	1244	1791	1543	1071	1350	1589	1279	1659	1491
6	1600	1488	1620	1526	1295	1028	1419	1367	1454	1447
7	1523	1317	1530	1374	1420	1305	1076	1286	1459	1305
8	1469	1414	1490	1422	1366	1407	1430	967	1535	1459
9	1624	1551	1475	1517	1556	1551	1403	1464	1077	1419
10	1638	1334	1558	1554	1461	1457	1273	1336	1386	1093

Table 1: This table shows the distances between the images of the first series (vertical entries) and of the second series (horizontal entries).

Fig. 8 illustrates graphically the results. A smaller distance is displayed as a darker pixel.

The distance function that describes the distance of one feature vector to all other feature vectors in the image has only a few evanescent minima if we use the modulus of the complex response as a distance measure. This leads to less reliable grid matchings than with the Gabor phase. Furthermore, Gabor phase produces more smoothly deformed grids than Gabor magnitude, and without additional smoothness constraints.

However, phase as criterion for comparing feature vectors results in a face recognition method that is not robust to large scaling. The distance function for feature vectors based on phase has a large number of well-defined local minima. The initial rigid translation of the grid prevents feature vectors to fall into wrong local minima, but a large scaling of the face in the image to be tested can cause the feature vectors to “drop” into the false minima.

Other features like local symmetry, [4], local orientation, [3], or orthogonal wavelets, may also be used as features. Especially if no rotation has to be expected from the pattern being sought, a fast wavelet transform could be used to accelerate the grid matching.

An image can be reconstructed partly from the feature vector graph to demonstrate the information contents of the labeled grid [25].

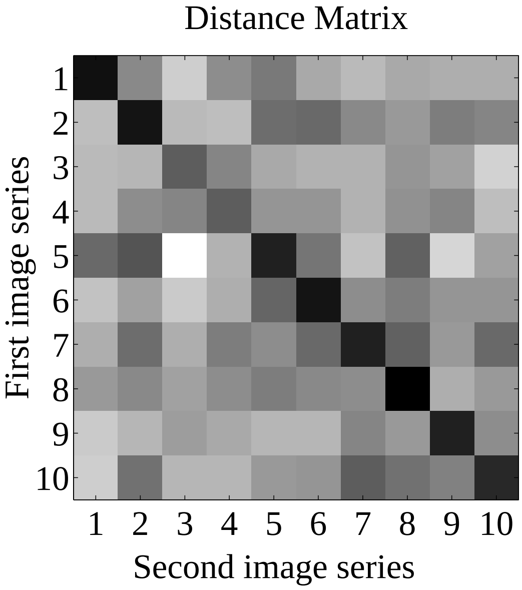


Figure 8: This matrix shows the face distances as grey levels. Two series of 10 images taken at different times have been used.

3.3 Structured light

We see structured light as a means to get rid of problems related to illumination and posture (orientation, scale). Furthermore much static (robust), discriminant and easy to detect characteristics lie in volume information (forehead, nose, cheek, lips, ...).

Structured light, [15], should offer a sufficient precision, remain simple and cheap, and be quick enough to avoid head motion defects and allow time integration. The 3D measurements are derived from triangulation thanks to a camera and a projector casting a pattern of lines. The technique allows to reconstruct sufficiently well behaved surfaces (such as a face) with only one image. Considering sequences of such images is a possible means to increase precision or enlarge the volume seen. The illumination disturbance (proper to active 3D systems) can be reduced by operating in near infrared.

Our preliminary tests with an off-the-shelf camera and a low cost projector have proved promising. We are currently developing an automatic calibration procedure to improve the precision further. (see Figure 9 and Figure 10). The characteristics brought by the 3D analysis will concern curvature, distance and volume of specific head parts. First, the head will be roughly analyzed in terms of curvature (convex, concave) to normalize the data with respect to orientation and scale variations (see [13]). Then curvature values, volume and distance between head parts will be extracted as features. This will bring new information in comparison with grey-level based features.

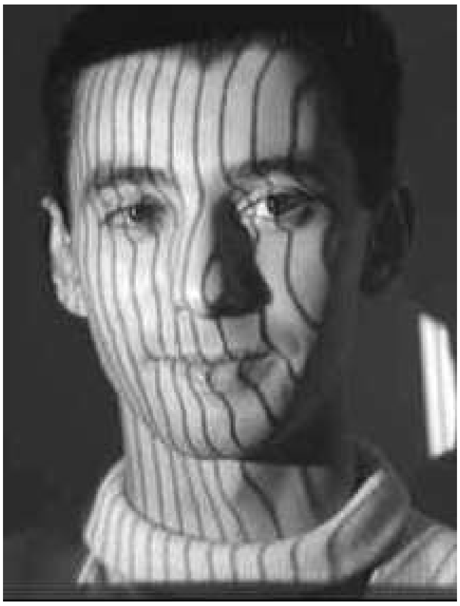


Figure 9: Image with structured light

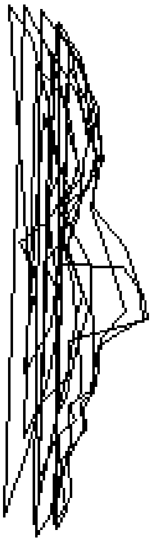


Figure 10: 3-D reconstruction

4 Verification by speech

4.1 System overview

The verification by speech is based on three methods for comparing the speech recorded for verification with a reference built from training utterances of the claimed person. These methods are: Dynamic Time Warping (DTW), Sphericity (SOSM), and Hidden Markov Models (HMM). Each method outputs a scalar value, which can be considered as a measure of conformity of the newly recorded speech with the stored speaker reference. Two of the methods (DTW and HMM) are specific to text-dependent speaker verification, i.e. to the case where the person says the same text when his identity is verified as he said during training. The third one (SOSM) can be used for both, text-dependent and text-independent, speaker verification. All the methods are used here in a text-dependent case.

The comparison of the speech recorded for verification with the speaker reference is performed on speech parameters which are extracted from the captured audio signal. The speech parameters we use are the Linear Prediction Cepstral Coefficients (LPCC) [1, 23], the aim of which is to model the vocal tract of the speaker. They are computed every 10ms, over a time window of 25ms. The way the speaker reference is built from the speech parameters of the training utterances depends on the comparison method (see section 4.2).

The verification decision, i.e. the decision whether to accept or reject the person, is taken based on the results of the comparison methods (conformity measures). In the case of a single comparison method, the decision is simply taken by thresholding the comparison result, the problem being to find the optimal threshold value. We have investigated different ways of setting the threshold (see section 4.3). In the case where the methods are combined, there are many possibilities for achieving a decision based on the comparison results. This is a problem going into the topic of modalities fusion, which is discussed in section 5.

4.2 Individual speech comparison methods

DTW, SOSM, and HMM are methods which have also been used in other speaker verification systems. We explain here only what is particular to our system. For this, we introduce the following definitions. We call training vector, the time sequence of LPC Coefficients extracted from the training speech, and test vector, the time sequence of LPCCs extracted from the speech recorded for verification, where both speech sequences are supposed to express the same text.

Dynamic Time Warping (DTW) A speaker reference vector is built in averaging the training vectors (of the same text). The dynamic comparison performed by the DTW method is applied between the test and reference vectors [14].

Sphericity (SOSM) A speaker reference vector is built in concatenating the training vectors (of the same text). The conformity measure is the sphericity measure computed from the covariance matrices of the reference and test vectors [9, 5].

Hidden Markov Models (HMM) We use left-right HMM structures [22]. The number of states per model has been determined empirically. There is approximatively one state for each phoneme and one state for each transition between phonemes. Each state contains a single Gaussian.

For each text unit (e.g. digit), there exists one model for each speaker and a world model. The two model types of the same text unit have the same structure. A likelihood ratio (the likelihood of a speaker model divided by the likelihood of the world model) is computed.

The world model is created from a database with a large number of speakers. The speaker model is initialized with the world model and re-estimated with the training vectors of the concerned speaker.

4.3 Acceptance/rejection decision

In case of a single comparison method (DTW, SOSM, or HMM), we have investigated three ways of setting the decision threshold: EER global, EER individual, and Furui [16].

Basically, the threshold value can be either general (i.e. the same for all the persons) or individual (personal). The latter solution has more flexibility and should give better results. However, it has to face the short-coming of individual training data to give a good estimate of the optimal threshold. "EER global" is a threshold common to all the persons. "EER individual" and "Furui" are person-dependent thresholds.

On the database used for experimentation (see section 4.4), the best results were obtained with the "Furui" threshold, for the DTW and SOSM methods, and with the "EER global", for the HMM method. This last result can be explained by the fact that, in the HMM case and not in the other ones, the comparison measure has been normalized by taking the ratio between the likelihood of the person model and the likehood of the world model.

In the case of multiple comparison methods, among the numerous possibilities for achieving a decision based on the comparison results, we have used the following procedure. Individual decisions are taken in applying a threshold on the individual comparison results and the global decision is realized in applying a majority law to the individual results. The performance achieved in combining the comparison methods in the way just described is presented in the next section.

4.4 Results

The results which are available now concern a set of 10 speakers, whose voice data is part of the Polycode database [20]. This database was recorded through a telephone line in several sessions. In each session, each speaker had to say, among other sentences in French, 5 times his 7-digits personal identification number (PIN) and 4 times all the 10 digits, in 4 different orders (which are the same for all the speakers).

The database has been divided into three disjoint sets of sessions: one for estimating method parameters, one for training the speaker references, and one for verification tests.

The data used for verification are as follows: for each speaker, 20 samples of his PIN number pronounced by himself and 9x20 samples of his PIN number constructed from 10-digits sequences pronounced by each of the other speakers. So, in total, there are 200 correct access trials and 1800 impostor trials.

The verification performance are expressed in table 2 by the False Acceptance rate (ratio of impostor trials falsely accepted) and the False Rejection rate (ratio of correct access trials falsely rejected).

Table 2 shows that combining the methods in the way described in the last section yelds an improvement of the false rejection rate. The false acceptance is slighthly deteriorated. This can be due to the small size of the database, as well as the crudeness of the fusion rule.

Method	FR% (200 tests)	FA% (1800 tests)
DTW / Furui threshold	23.5	7.67
SOSM / Furui threshold	14.0	5.28
HMM / EER global threshold	3.5	2.5
Combined Decision	2.0	2.72

Table 2: Speech-based verification performance (individual and combined methods)

5 Fusion strategies

There are many possible ways of combining modalities. In principle, the fusion of modalities could take place at any step of the "classical" pattern recognition processing chain: *data*, *features*, or *decision fusion* [8]. However, since the methods we have developed so far for features extraction and comparison are different from modality to modality, the fusion has to occur in the last stage, i.e. the fusion is concerned with taking the acceptance/rejection decision based on the output given by the different comparison methods.

Up to now, fusion has only been studied and experimented with respect to the three voice recognition methods. The problem of combining the comparison outputs for taking a decision has been solved in applying to each individual comparison result a proper decision threshold and in combining the individual decisions with a majority law (see section 4.3).

We are currently extending the fusion to the visual features and we are investigating other ways of solving the problem, e.g. we are considering the possibility of weighting the individual decision outputs. Rather than using constant weights, we are investigating the possibilities of adapting the weights to a person or a person-class, or to particular conditions where a method is more or less powerful.

6 Conclusion

Concerning visual aspects, we have shown that a technique combining Gabor phase information with dynamic link matching is able to discriminate faces. The structured light experiments have shown that 3-D biometric information is cheaply accessible. Further studies to improve its precision is underway. Biometric features concerning the facial features in terms of inter distances of eyes, mouth and nose necessitate a robust localization of these features. Our tests indicate that model (such as ellipsoids) based fitting is a way to improve the robustness. Other easily available data from video sequences, such as profiles [2], could be used to enhance the capacity of the image based verification.

With respect to voice, we have shown that 3 different recognition methods can individually achieve promising performance.

The combination of voice recognition methods achieves an increase of performances, letting augur good results for the fusion with face features we are currently realizing. Also foreseen is the test or estimation of performance achievable of a large multi-modal database.

References

[1] B. S. Atal. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". *JASA*, Vol. 55, No. 6, pp. 1304–1312, (1974).

[2] C. Beumier, M.P. Acheroy, "Automatic Face Identification", In *Applications of Digital Image Processing XVIII*, SPIE, vol. 2564, pp. 311-323, July (1995)

[3] J. Bigun, G. H. Granlund and J. Wiklund *Multidimensional orientation estimation with applications to texture analysis and optical flow* IEEE-PAMI vol. 13, No. 8, pp. 775-790 (1991).

[4] J. Bigun *A structure feature for image processing applications based on spiral functions* Computer vision, graphics and image processing. No. 51, pp. 166-194 (1990)

[5] F. Bimbot and L. Mathan. "Second-order statistical measures for text-independant speaker identification". In ESCA [10], pp. 51–54.

[6] R. Brunelli and D. Falavigna. "Person identification using multiple cues". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 10, pp. 955–966, October (1995).

[7] Y. Dai, Y. Nakano, "Extraction of Facial Images from Complex Background Using Color Information and SGLD Matrices," *Int. Workshop on Automatic Face- and Gesture Recognition*, pp. 238-242, ed. Martin Bichsel, Zurich, Switzerland, June (1995).

[8] Belur V. Dasarathy. *Decision Fusion.* IEEE Computer Society Press, Los Alamitos, California, 1994.

[9] K. Drouiche. *Chapitre IV: Test de sphéricité.* PhD thesis, ENST, Jan. (1993).

[10] ESCA, editor. *ESCA Workshop on Automatic Speaker Recognition Identification Verification.* ESCA, April (1994).

[11] A. Eleftheriadis, A. Jacquin, "Automatic face location and tracking for model-assisted coding of video teleconferencing sequences at low bit rates," *Signal Processing: Image Communication*, vol. 7, no. 3, pp. 231-248, July (1995).

[12] G. Galicia, A. Zakhor, "Depth Based Recovery of Human Facial Features from Video Sequences," *IEEE Conf. on Image Processing*, vol. 2, pp. 603-606, Washington, USA, October (1995).

[13] G.G. Gordon, "Face recognition based on depth maps and surface curvature", In *Geometric methods in Computer Vision*, SPIE, vol 1570, San Diego (1991).

[14] M. Homayounpour. *Vérification du locuteur: Dépendante et indépendante du texte.* PhD thesis, Université PARIS-SUD, (1995).

[15] R.A. Jarvis, "A perspective on range finding techniques for computer vision", In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, pp 122-139, March (1983).

[16] D. Genoud, G. Gravier, F. Bimbot, M. Homayounpour, and G. Chollet. Amélioration des performances de reconnaissance du locuteur par combinaison de méthodes. accepted for XXIèmes JOURNÉES D'ÉTUDES SUR LA PAROLE, Avignon 10-14 Juin 1996.

[17] M. Lades, J. Buhmann J. C. Vorbrüggen, J. Lange, C. v.d. Malsburg, R. P. Würtz, and W. Konen. "Distortion invariant object recognition in the dynamic link architecture.". *IEEE Transactions on Computers*, Vol. 42, No. 3, pp. 300–311, March (1993).

[18] H. Niemann, "Pattern Analysis and Understanding", Springer-Verlag, (1990).

[19] I. Pitas, A. N. Venetsanopoulos, "Nonlinear Digital Filters: Principles and Applications", Kluwer Academic Publishers, (1990).

[20] Dominique Genoud and Gérard Chollet. Polycode a verification database. Technical report, IDIAP, CH-1920 Martigny, 1995.

[21] D. A. Reynolds. *A Gaussian mixture modeling approach to text-independent speaker identification.* PhD thesis, Georgia Institute of Technology, (1992).

[22] A.E. Rosenberg & C.H. Lee & S. Gokoen. Connected word talker verification using whole word hidden markov model. In *ICASSP-91*, pages 381–384, 1991.

[23] P. Thevenaz. *Résidu de prédiction linéaire et reconnaissance de locuteurs indépendante du texte.* PhD thesis, Université de Neuchâtel, (1993).

[24] H. Wu, Q. Chen, M. Yachida, "An Application of Fuzzy Theory: Face Detection," *Int. Workshop on Automatic Face- and Gesture Recognition*, pp. 314-319, ed. Martin Bichsel, Zurich, Switzerland, June (1995).

[25] R. P. Würtz. *Multilayer Dynamic Link Networks for Establishing Image Point Correspondences and Visual Object Recognition*, Vol. 41 of *Reihe Physik*. Verlag Harri Deutsch, Thun, Frankfurt am Main, (1995).