Person Recognition Based On Lip Movements

Olga Shipilova

Laboratory of Appl. Mathematics, Department of Information Technology
Lappeenranta University of Technology
olga.shipilova@lut.fi

Abstract. The paper presents a person recognition problem using static and dynamic features of a moth area. The basic concept of construction of the *Lip Model* to localize and track lip on an image sequence is provided. Two probability technologies of recognition implementation, based on HMMs and GMMs modeling, are described. Applications of lip motion in recognition systems are presented.

Introduction

Person verification plays a grate role in everyday life. Many factors of modern society enforce to increase accuracy and robustness of existing verification methods such as password, photo identity card, Personal Identification Number (PIN) code and keys, which can be inadequate to meet heavy demands. In such cases technologies based on biometrics become more attractive.

Def. Biometric is a physical or behavioral person characteristic, which is universal, unique, permanent and collectable [4].

Basically, recognition systems deal with identification and authentication. Identification or so-called 1:N matching means the determination of the corresponding person from a database containing many people, or decision that the person is not enrolled in the database. In authentication or 1:1 matching the input biometric is matched against a single biometric record [10].

There are many person characteristics, which have been evaluated for identification and authentication systems. Most commonly used biometrics such as fingerprint, face, voice, iris, signature, retinal scans, hand and finger geometry are physical characteristics, whereas signature and keystroke dynamics are behavioral ones. All these characteristics have advantages and disadvantages. For example, one can readily cite factories, which influence the performance of voice-based and face-based recognition systems.

Voice-based recognition is highly dependent on the following factors [7]:

- Channel distinction between training and test set
- Environmental and channel noise
- Speaker inconstancy due to speaking rate, speaking level
- Time interval between training and test

- Time interval of test and training

Face-based system is mainly affected by the following conditions [7]:

- Illumination
- Orientation of face in 3D space
- Speaking and facial expression
- Hairstyle, glasses, makeup and etc.

The problem of such impact on speechriding and on face recognition has led to the technology for person identification and authentication, rested upon spatial and temporal analysis of image sequences of the talking face, in the other words, lip movements.

First of all lip motion presents speech dependent information. But the static and dynamic features of the lips also contain important speaker dependent information. [11]. The ability of labial movements to transmit information in speech appreciation has been studied extensively and the visual signal has been successfully exploited in speech recognition systems. The fact that temporal lip information not only presents speech information but also characteristic information about a person's identity has been neglected until 1996, when the group of Luettin (University of Sheffield, UK) has suggested a new modality for person verification or recognition based on "spatio-temporal lip features" [7]. Mainly Luettin's group has investigated the problem of speech recognition. In the range of this research they have been interested how lip movements could be used for person identification or authentication. Luettin's research has been continued and extended in the many investigations of such scientists as C.C Broun and X Zhang [2], T. Wark and D Thambiratnam [10] and many others.

Biometric Recognition System

Scientists of IBM Thomas J. Watson Research Center have suggested considering a recognition system based on biometrics as a generic pattern recognition system shown in Fig. 1 [10].

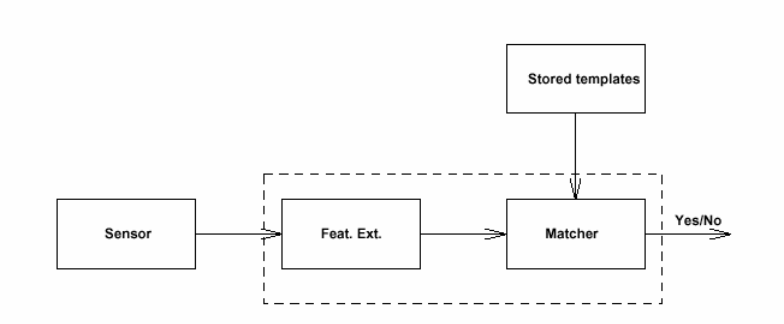

Fig. 1. A generic biometrics system.

The generic biometrics system consists of four parts. The *input* subsystem involves the special sensors needed to acquire the biometric signal. The original signal consists

of both required identifying information and irrelevant information. In the next stage, the *feature extraction* part or subsystem picks up invariant features from the signal and forms reference or input templates. The recognition system possesses the representation basic templates. Further the *matching* subsystem parallels the reference and basic templates and returns the degree of match or mismatch as a score. Finally, this score is compared with a decision threshold to define the comparison is matched or it is not matched. Efficiency and performance of the full recognition system seem to depend on efficiency and performance of all the subsystems. Additionally, the system has to have efficient storage and retrieval, error free transmission, moreover, methods for encryption and decryption of the result must be provided [10].

The performance of the biometric system is assessed by a hypothesis testing framework. It is assumed the basic biometric template is pattern [10].

$$P' = S(B'),$$

and the input one is pattern

$$P = S(B).$$

In the range of the identity purpose null and alternative hypotheses are considered [10].

$$H_0 : B = B', \text{ the required identity is correct;}$$

$$H_1 : B \neq B', \text{ the required identity is } \textit{not} \text{ correct.}$$

In the terms of a score $s = \dim(P, P')$ and a decision threshold T_d null and alternative hypotheses are represented as [10].

$$H_0 : s \geq T_d, \text{ the required identity is correct;}$$

$$H_1 : s < T_d, \text{ the required identity is } \textit{not} \text{ correct.}$$

To evaluate accuracy of the recognition system for the given decision threshold two main quantities are used, namely the False Accept Rate (FAR) and the False Reject Rate (FRR).

Def. FAR is proportion of *non-mated* pairs resulting in *false acceptance* (deciding H_0, when H_1 is true) [10].

Def. FRR is proportion of *mated* pairs resulting in *false rejection* (deciding H_1, when H_0 is true) [10].

The ratio between FAR and FRR is depicted in Fig. 2, where ***p*** is probability [10].

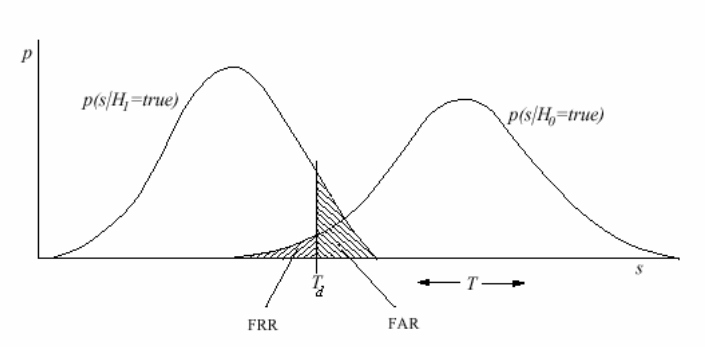

Fig. 2. False (left) and true (right) distributions with classification error definitions.

By varying the FAR and the FRR, the Receiver Operating Characteristic (ROC) curve is obtained, Fig. 3 [10].

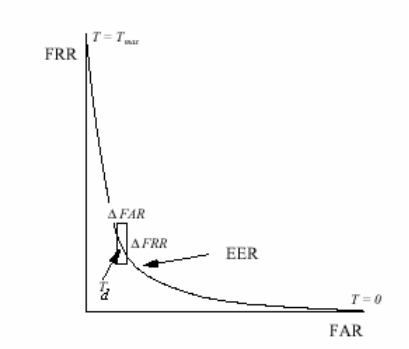

Fig. 3. Receiver Operating Curve (ROC).

A scalar figure of merit used to judge the performance of a recognition algorithm, is the so-called Equal Error Rate (EER), corresponding to the ROC operating point having FAR=FRR, Fig 3 [10].

Person Identification and Verification

Person recognition tasks can be divided into two classes:

- Person identification
- Person verification

The task of the first class is to determine the person from a closed set, whose features best match the features of the person to identity. Thus, it is assumed that only enrolled person will access the identification system [7]. In Fig. 4 a diagram of person identification is depicted.

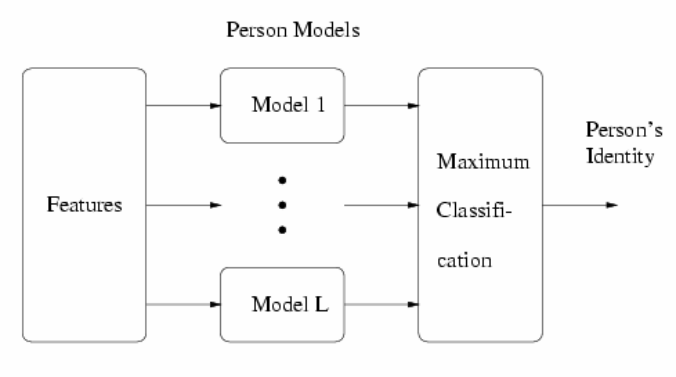

Fig. 4. Person identification system [7].

The verification problem is to validate the claimed identity of a person from an open set. Therefore, verification system has a reject subjects, called by *impostor* [7]. Fig. 5 demonstrates the scheme of a person verification system. Verification is basically executed by the comparison the claimed client model with the impostor model.

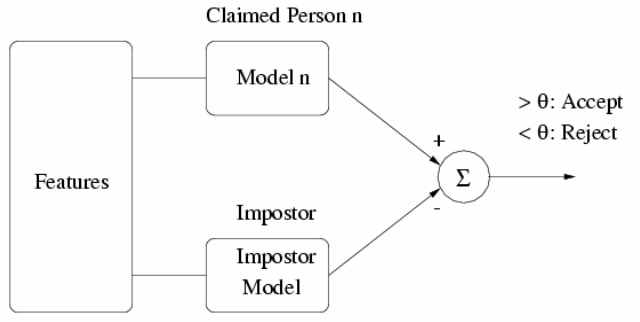

Fig. 5. Person verification system [7].

A problem of identification or authentication based on lip movements can be derived into four parts or subproblems:

- Defining of Region of Interest (ROI)
- Constraining Lip Model from the training set
- Lip tracking
- Implementing probability analysis

At the beginning the recognition system has to extract picture of lips from a video sequence. Then, main features of the static and moving lips are defined and a lip model is constructed. By using the lip model localization and tracking of lip on the image sequence is implemented. Finally, probability analysis, rested upon the chosen database, is performed and judgment whether the person is recognized is pronounced.

Lip Model

Region of interest

Person recognition using lip information is started with obtaining a so-called Region of Interest (ROI). Further the recognition system works with the ROI. There are two types of database which can be used for person recognition based on the labial area. The first kind of databases consists of face images. In this case the recognition system has to be able to localize the lip area in the image. This is commonly executed by an assumption that the distance between the outer points of eyes is equal, approximately, to the distance between the bridge of noise and the middle of the mouth area [11], Fig. 6.

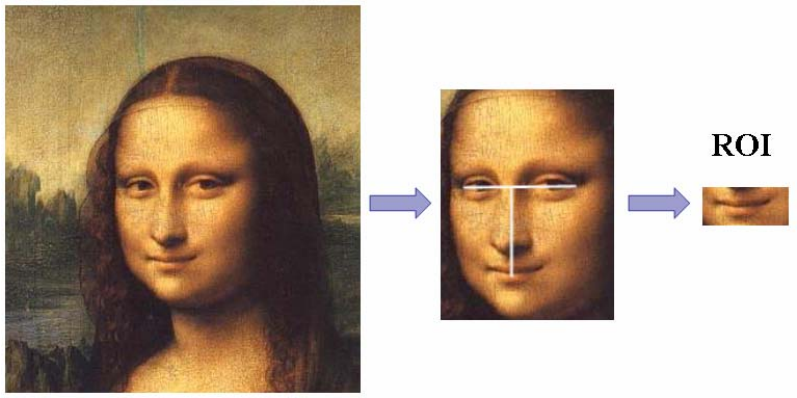

Fig. 6. Region of Interest (ROI)

In the second kind of databases video images are confined to a square area around the mouth, thus the ROI is already known, Fig. 7.

Fig. 7. Tulips1 database [7].

Having defined the ROI, the system has to make an appropriate description of visible motion. Since a physical process is modeled, it could be described in terms of physical movements and positions of the articulators, for instance, the muscle action could be assessed. However, the motion of mouth region musculature is very complex because of this motion is three-dimensional, not directly observed and there are at least thirteen groups of muscles defining lip movements [7].

This chapter describes the Luettin's method for visual features extraction to be used for speaker recognition applications [6], [7], [8]. The technology rests upon

construction of Appearance Based Model (ABM) for lip localization, lip tracking and features extraction. This kind of models involves static lips information, namely the inner and outer lip counters. During lip extraction, the Appearance Based Model is expanded to the Active Shape Model (ASM), based on the Point Distribution Model (PDM). The ASM is used to represent shape deformation and consists of two parts - the shape submodel and the intensity submodel.

Fig. 8 and 9 show diagrams of the different stages for training and using the lip model.

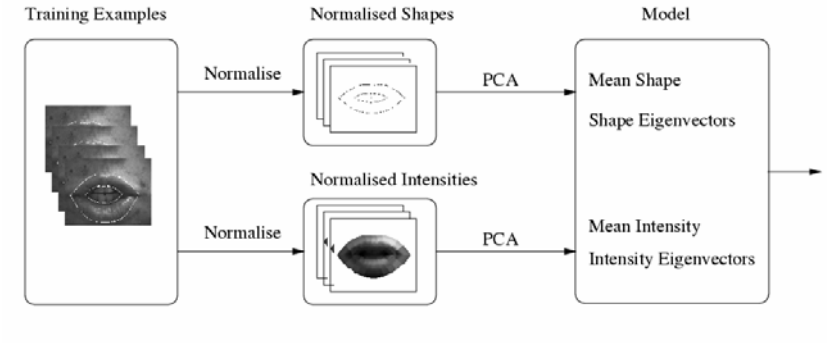

Fig. 8. Training the lip model [7].

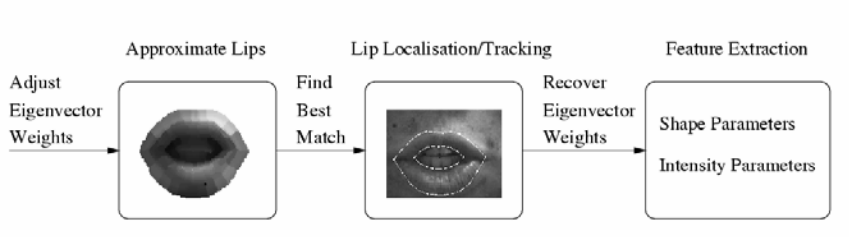

Fig. 9. Using the lip model [7].

Shape submodel

The first step of shape modeling is to label training shapes consisting of points around the bound of the training examples. To place this points collection and to compare equivalent points from different shapes the reference pointes are needed. In the work of Luettin the *Model DC* is described where the two outer points of the lip are used as reference system. Their distance is scale, their orientation with respect to the horizontal is defined as the angel, and center of the scale as the origin. The other points are fixed at equal horizontal distance along the lip contour. In the Fig. 10 Model DC with translation t_x and t_y, scale s and angle θ is depicted [8].

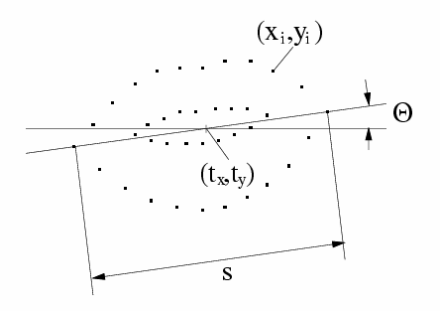

Fig. 10. Model DC [8].

Each shape of the training set is represented as a vector of coordinates of each point. The training shapes are then normalized to connect different samples without distortion from scale, translation or rotation.

Thus for a given set of normalized labeled shapes, the mean shape $\bar{\mathbf{x}}$ and the covariance matrix $\mathbf{S}_s$ can be calculated. The eigenvectors and eigenvalues of the matrix $\mathbf{S}_s$ are obtained by Principal Component Analysis (PCA). The eigenvectors with the largest eigenvalues mean the most important modes of variation.

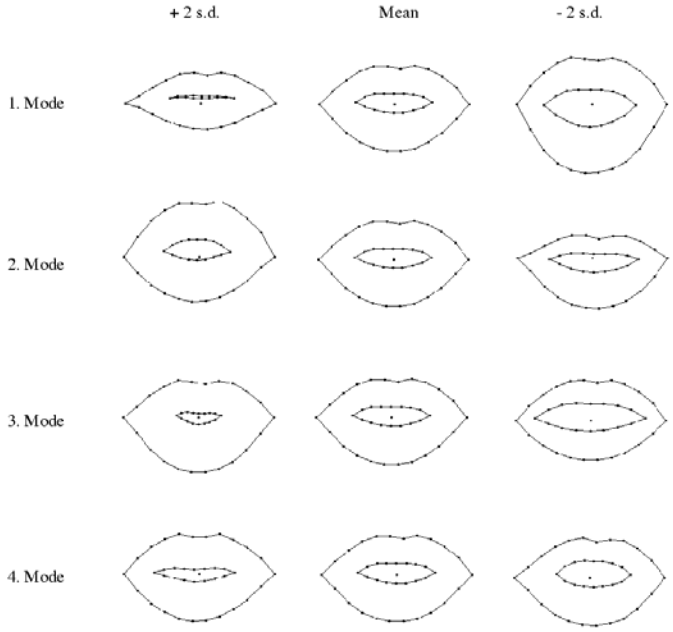

Fig. 11. Shape variability for Model CD [7].

Fig. 11 shows an example of modeling shape variability, where the middle column displays the mean shape for Model DC and right and left ones demonstrate shape variation as the sum of the mean shape and different basis shapes.

Intensity submodel

To define and then to extract significant features of lip contours for speaker recognition the intensity submodel is used.

The most common techniques to obtain the contours are based on using edges or gradients. However the technologies have a number of disadvantages. So, the edges description of lips bounds is ill-posed and does not have unique solution. For instance, the semblance of lip contours is very changeable even for the same person, Fig. 12.

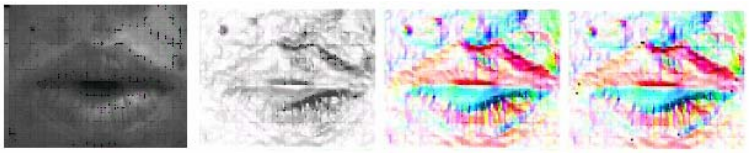

Fig. 12. Contour extraction by using edge technology [5].

In turn the inner contour gradient depends on mouth, which can be opened or not, and teeth and tongue, which can be visible or not. The speaker features, such as make up, facial hair and ethnic origin influence gradients, Fig. 13. Moreover environmental conditions, for example, illumination, has an impact on value of gradients, Fig. 13.

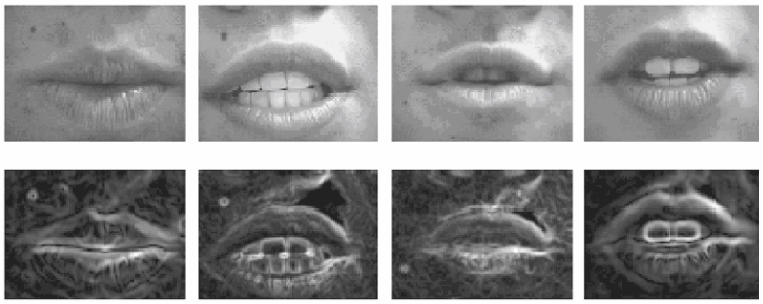

Fig. 13. Contour extraction by using gradients technology [7].

One more appropriate way to reach main lip contours features is to investigate grey-level appearance both at each contour point and in its vicinity. The grey-level vectors are perpendicular to the lop bounder and centered at the model point. The statistics of the current grey-level semblance are caught around each model point and their dominant variation modes are evaluated from *training set*. To obtain grey-level vector is demonstrated in Fig. 14.

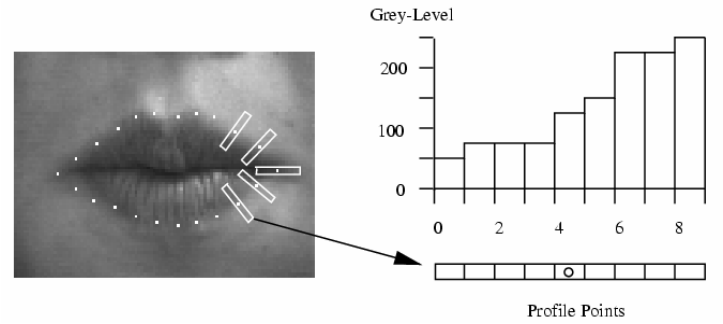

Fig. 14. Extraction of grey-level sample [7].

In same way as for shape modeling the main eigenvectors corresponding the largest eigenvalues are calculated by means of PCA. The class of such kind of eigenvectors and its eigenvalues represents the intensity around the lip contour and takes into account properties like, for instance, the intensity of visibility of tongue and teeth, the oral cavity, protuberance. Therefore these parameters are dominant features with respect to person authentication. Fig. 15 shows an example of the intensity submodel, where the middle column displays the mean intensity and right and left ones demonstrate the three most significant modes of intensity variation.

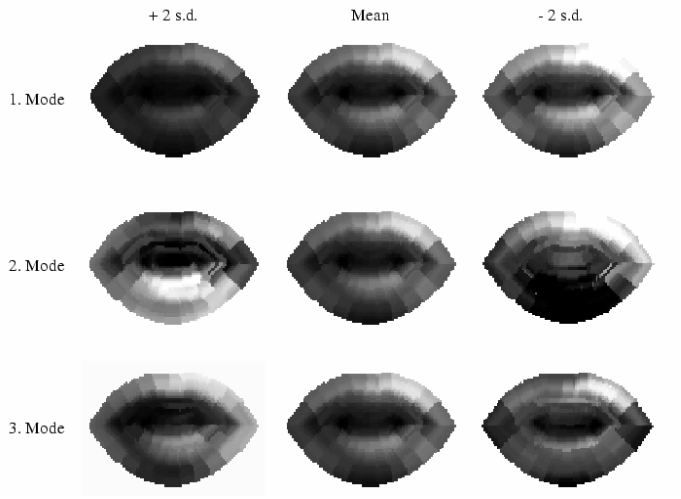

Fig. 15. Intensity variability of Model CD [7].

Thus the grey-level model describes the area overplayed by the profile vectors, videlicet the lip area, the mouth opening or closing and the skin around the lips. Such a sample extracts the global variation across different person (speaker, in the case of lip verification) and the variation within one person.

Having combined the shape and intensity sub-models, the lip model Ω can be represented as

$$\Omega = \left(\bar{\mathbf{h}}, \bar{\mathbf{x}}, b_s, b_i, t_x, t_y, s, \theta\right).$$

This definition consists of all parameters to approximate the shape, intensity and pose of lip image, there are $\bar{\mathbf{h}}$ is global main intensity profile, $\bar{\mathbf{x}}$ is normalized training shapes, b_s is a vector of weights of main eigenvectors for the shape model, b_i is a vector of weights of main eigenvectors for the intensity model, t_x and t_y are coordinates of the origin for Model DC, s is scale for Model DC and θ is the angel orientation with respect to the horizontal [7].

Lip tracking

The next subproblem of feature extraction is to mate the lip model obtained and lip image given. In other words lip tracking is the problem of localizing the lips in the image. This is commonly executed by matching the model to the image at different places and with different shapes and by choosing the hypothesis with the highest probability. To describe the fit between the model and the image there are many different cost functions, for instance, the image-gradient function and the intensity based function [7].

Probability Analysis for Person Recognition

The lip model described above is used to locate and track the lips of talking person and to extract speaker dependent information from the motion sequence. Models of speaking persons are constructed from a training set of these extracted features. Person recognition is implemented by capturing the same spatio-temporal features the test person and by choosing that person, whose model has the highest probability as the identified person [6].

Two different modeling techniques can be suggested for purposes of person identification or verification. One is text-dependent and the other is text-independent. For the first approach the speech used to train and test the system must be the same, while for the second one it is not important.

The approach based on Hidden Markov Speaker Models (HMMs) with the mixtures of Gaussian distribution is text-dependent one. In the range of this technology a speaker (recognized person) is denoted by a set of HMMs, constructing the set of speech groups spoken by that person. The speech groups may be phonemes or words [7]. Each subject and each speech group is represented by one model built. Identification is executed by maximum of a posteriori probability based on the representation of the person model and the observation sequence.

The example of text-independent approach is one rested upon Gaussian Mixture Speaker Models. For this technology it is assumed that the mixture components describe a basic set of probable lip states which are characteristic for a certain person. Each speaker is defined by only one model describing all utterances of that person.

Person recognition for the method is executed by estimating the posterior probability of built model from the observed sequence of person features [7].

Schematically, these two methods are demonstrated in Fig. 13.

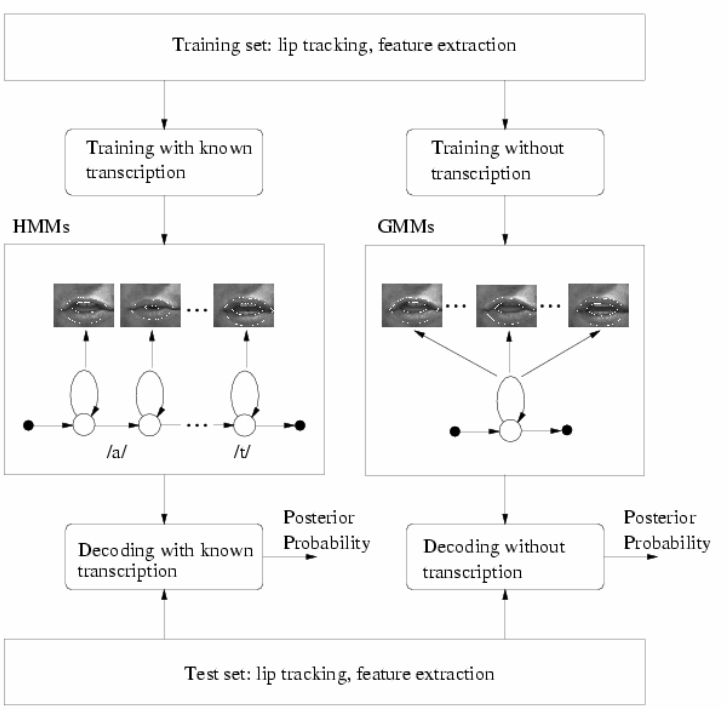

Fig. 16. Methods for visual speaker recognition based on HMMs (right) and GMMs (left) [7].

Experiments

The Luettin's approach of constructing the lip model and lip localization and tracking has been tested on the audio-visual Tulips 1 database. It involves 96 grey-level images sequences of 12 speakers. The images are confined to the mouth area of the speakers. The accuracy for text-dependent (TD) and text-independent (TI) person identification is demonstrated in Table 1 [6].

	Shape	Intensity	Shape + Intensity
TD	72.9 %	89.6%	91.7%
TI	83.3%	95.8%	97.9%

Table 2. Person identication tests

Best performance has been reached by using lip model, mixing shape and intensity simulations.

Applications

Lip motion as biometric provides information which can be easily combined with face and voice information [4]. This ability is strongly used to develop multimodal recognition systems based on two or more biometrics. The examples of such projects are M2VTS and SESAM.

M2VTS

The method described above is used in the range of M2VTS project. This is a multimodal person recognition system for teleservice and security programs [1]. The purpose is to give a full solution of security access. The M2VTS system is based on two cues that are voice record and video sequence of tacking face [9].

SESAM

The person authentication system SESAM contains three different biometrics from two different data origins: one static signal gained from an image of the face and two dynamic signals, the spectrum of voice and the lip motion of a person saying its name in front of system [4].

	Identification	Verification
Classification	1-FAR %	1-FAR %
Speech	89.6	98.1
Lip motion	89.0	96.5
Face	81.3	97.3
Combination	93.0	99.7

Table 3. Test results for SESAM

The SESAM system has been tested in Fraunhofer-Institute for Integrated Circuits, Germany. A spatial recoding station, consisting of a standart camera and a microphone, has been developed to cath optical and acoustical data. The experiments

have shown good performance of the system both for identication and verification problem. The resultes of the test are presented in Table 2 [3].

Conclusions

In this paper a person recognition problem based on lip movements has been presented. The general technology for lip-based recognition system, suggested by the Luitten's group has been considered [7].

A problem of identification or authentication based on lip movements can be derived into four parts or subproblems. Person recognition is started with obtaining the so-called Region of Interest (ROI), which confining within the mouth area [11]. Then, main features of the static and moving lip are defined and a lip model is constructed. The model rests upon shape and intensity subparts. The shape submodel involves static lips information, namely the inner and outer lip counters, and is based on the Point Distribution Model (PDM). The intensity submodel is used to define and then to extract significant features of lip contours. To reach main lip contours features grey-level appearance both at each contour point and in its vicinity is investigated [7].

Having normalized the shape and intensity forms, the lip model is applied to localize and to track the lips in the image sequence. This is commonly executed by matching the model to the image at different places and with different shapes and by choosing the hypothesis with the highest probability. To describe the fit between the model and the image there are many different cost functions such as the image-gradient function and the intensity based function [7].

Finally, probability analysis, based on the chosen database, is performed and judgment whether the person is recognized is pronounced. Two different modeling techniques can be suggested for purposes of person identification or verification. The approach based on Hidden Markov Speaker Models (HMMs) with the mixtures of Gaussian distribution is text-dependent one. For this technology a speaker is described by a set of HMMs, constructing the set of speech groups spoken by that person. The speech groups may be phonemes or words [7]. The text-independent technology is rested upon Gaussian Mixture Speaker Models. For this technology each speaker is defined by only one model describing all utterances of that person. Identification in the range of both approaches is executed by evaluation of maximum a posteriori probability [7].

The experiments, carried out by the Luittin's group, have shown better results of lip-based recognition system than face-based or voice-based ones.

It must be emphasized lip motion as biometric provides information which can be easily combined with face and voice information [4]. This ability is extensively used to develop multimodal recognition systems based on two or more biometrics. The examples of these projects are M2VTS and SESAM [1], [3].

References

1. Archeroy M. et al. "*Multi-modal Person Vetrification Tools Using Speech and Images*", Proc. Europ. Conf. on Multimedia Applications, Services and Techniques, 1996
2. Broun C.C., Zhang X., Mersereau R.M., Clements M. "*Automatic Speechreading with Application to Speaker Verification*", IEEE, 2002
3. Dieckmann U., Plankensteiner P., Wagner T. "*SESAM: a Biometric Person Identication System using Sensor Fusion*", Pattern Recognition Letters, 18, 1997, p.827-833.
4. Dugelay J.-L., Junqua J.-C., Kotropoulus C., Kuhn R., Perronin R., Pitas I. "*Recent Advances in Biometric Person Authentication*", http://www.eurecom.fr/~perronni/papers/icassp02.pdf, 15 Nov, 2003.
5. Gordan M., Kotropoulos C., Pitas I. "*Pseudoautomatic Lip Contour Detection Based on EdgeDirection Patterns*", http://www.muhci.org/papers/, 1 Nov, 2003
6. Jourlin P., Luettin J., Genoud D., Wassner H. "*Acoustic-Labial Speaker Verification*", Pattern Recognition Letters, 18, 1997, p.853-858
7. Luettin J. "*Visual Speech and Speaker Recognition*", PhD thesis, University of Sheffield, United Kingdom, 1997
8. Luettin J., Thacker N.A., Beet S.W. "*Active Shape Models for Visual Speech Feature Extraction*", http://www.idiap.ch/publications/luettin-nato96.bib.abs.html, 16 Nov, 2003
9. Luettin J., Thacker N.A., Beet S.W. "*Speaker Identification by Lipreading*", International Conference on Spoken Processing, USA, 1996
10. Ratha N.K., Senior A., Bolle R.M. "*Automated Biometrics*", http://www.research.ibm.com/ecvg/pubs/ratha-auto.pdf, 1 Nov, 2003.
11. Wark T., Thambiratnam D., Sridharan S. "*Person Authentication using Lip Information*", IEEE TENCON, 1997, p.153-156