Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 28 (2007) 1368–1382

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Audio–visual person authentication using lip-motion from orientation maps

Maycel-Isaac Faraj *, Josef Bigun

School of Information Science, Computer and Electrical Engineering (IDE), Halmstad University, Box 823, SE-301 18 Halmstad, Sweden

Available online 19 March 2007

Abstract

This paper describes a new identity authentication technique by a synergetic use of lip-motion and speech. The lip-motion is defined as the distribution of apparent velocities in the movement of brightness patterns in an image and is estimated by computing the velocity components of the structure tensor by 1D processing, in 2D manifolds. Since the velocities are computed without extracting the speaker's lip-contours, more robust visual features can be obtained in comparison to motion features extracted from lip-contours. The motion estimations are performed in a rectangular lip-region, which affords increased computational efficiency. A person authentication implementation based on lip-movements and speech is presented along with experiments exhibiting a recognition rate of 98%. Besides its value in authentication, the technique can be used naturally to evaluate the "liveness" of someone speaking as it can be used in text-prompted dialogue. The XM2VTS database was used for performance quantification as it is currently the largest publicly available database (≈300 persons) containing both lip-motion and speech. Comparisons with other techniques are presented.

Keywords: Audio–visual recognition; Biometrics; Biometric recognition; Speaker verification; Speaker authentication; Person identification; Lip-movements; Motion; Structure tensor; Orientation; Optical flow; Hidden Markov model; Gaussian Markov model

1. Introduction

The interactive recognition of a person represents an important challenge for automatic identity verification systems. Solution approaches include image sequence-based speaker verification systems, e.g. using linear discriminant analysis, LDA (Tang and Li, 2004), integration of audio and visual features into an automatic speech recognizer (Potamianos et al., 2003). An advantage of the interactive person recognition is thus its ability to prevent impostor attacks that use prerecorded facial images or speech data. It is more difficult to play a video sequence of a person if the system prompts the text to be spoken. The performance of multimodal systems using audio and visual information in biometrics are known to be superior to those of the acoustic and visual subsystems (Brunelli and Falavigna, 1995; Chibelushi et al., 2002; Tang and Li, 2001; Bigun et al., 1997).

Lip-movement analysis allows detecting changes in facial expression to recognize the spoken word. It can be used in speech recognition systems to provide complementary information (Yamamoto et al., 1998; Wark et al., 1999; Chen, 2001; Lucey et al., 2005), leading to improved speaker recognition performance as demonstrated by Jourlin et al. (1997), Luettin et al. (1996), Tang and Li (2001) and Hazen (2006). As lip-movements contain dynamic image information, they are very different from features extracted from audio and still images. Previous results (Jourlin et al., 1997; Dupont and Luettin, 2000), which employed lip-motion features extracted by tracking the lip-contours from still images in an image sequence of talking faces, reported noise resilience and improved recognition performance over acoustic and image based systems. Requiring manual interventions, however, lip-contours

[*] Corresponding author.
E-mail addresses: maycel.faraj@ide.hh.se (M.-I. Faraj), josef.bigun@ide.hh.se (J. Bigun).
URLs: http://www2.hh.se/staff/mafa (M.-I. Faraj), http://www2.hh.se/staff/josef (J. Bigun).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.02.017

extraction is, in itself, not robust to image-noise and hence requires high quality images. Another disadvantage is the fluctuating computation time due to the iterative convergence process of the contour-extraction.

This paper describes an algorithm that takes advantage of the low-level spatio-temporal information in an image sequence containing lip-motion. The motion in dynamic lip-images is modeled by moving-line patterns in space–time planes, where the normal of the space–time planes encode the normal velocity of the moving lines. We will present a normal velocity estimation method based on the 3D spatio-temporal space (Bigun et al., 1991) but using only 1D signal processing embedded in 2D manifolds (Faraj and Bigun, 2006; Kollreider et al., 2005). The lip-area is divided into several regions where motion statistics from predefined orientations are extracted for further use in person authentication. This results in increased computational efficiency compared to using the full 3D structure tensor for normal velocity estimations. Velocities are determined by combining two structure tensors where each tensor is computed in 2D by cascades of 1D filters. In the next section we discuss the normal velocity estimation model. In Section 3 we present an implementation of the algorithm along with normal velocity estimation results on synthetic test data, a rotating fan and expanding circles with different speed and spatial frequencies. In Section 4, we show usefulness of these results by suggesting novel lip-movement signatures for person authentication. Finally, we report on an experimental study using both lip-motion and speech features using the largest audio–visual database that is publicly available, along with comparisons in Section 5.

2. Velocity estimation by orientation detection

We describe our *normal image velocity* or *normal optical flow* estimation technique, which we later use to extract visual features for audio–visual, interactive person verification. A *dense optical flow*, can be determined by eigenvalue analysis of the structure tensor (Bigun et al., 1991). This method requires however multiple image frames since it simultaneously derives the velocity of moving points and lines. Accordingly, the computations can be excessive for applications that only need line motion features. Assuming that only line motion can be observed in local images, the computations can instead be carried out in 2D subspaces of the 3D spatio-temporal space. For lip-motion in image sequences, this assumption is realistic, as our experiments that we present later indicate.

Let $f(x,y,t)$ represent the intensity (gray-value) of a local image point (Fig. 1) in a spatio-temporal image where the parameters x and y represent the spatial coordinates, and t represents the time coordinate of the local image point. Here we assume that f is generated by a line translated in the normal direction with a certain velocity. The velocity component of translation parallel to the line is not possible to obtain; this is referred to as the *aperture problem*. Fig. 1 illustrates a space–time image of a moving lip represented by a sequence of 2D images. At any local image point we will study the orthogonal cross-sections as regards motion. This is illustrated by the white lines in the figure for a local image point (x,y,t). By cutting a cross-section through the cube, horizontally we can obtain an xt-slice marked by the horizontal white line (Fig. 1). Similarly, we can obtain a vertical cross-section, as marked by the vertical white line (Fig. 1). In the yt cross-section which can be regarded as a 2D space–time manifold, we can see the motion of the lip-sequence and mouth orientation in different modes. In the silent mode, we see that the local line directions in the yt-image will be horizontal, which corresponds to no motion in the 2D space–time. By contrast, in the speech mode, the local line-directions will be oblique. Next, we illustrate a local image containing a moving line in the xy-manifold which generates a plane in the xyt-space (Fig. 2). The observable velocity, which is the velocity in the normal direction of the line, is encoded by the orientation of the spatio-temporal plane in the xyt-space.

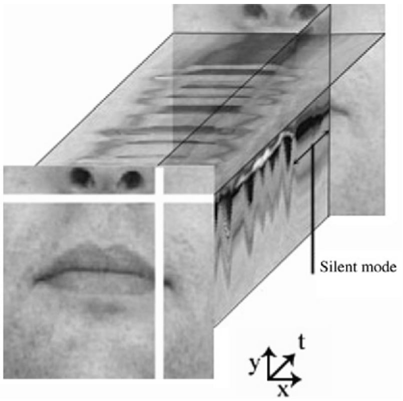


Fig. 1. A image sequence illustrated with a talking scene from the XM2VTS database. A yt-slice marked by the vertical white line and an xt-slice marked by the horizontal line show the surfaces in which local directions encode the local velocities. The horizontal directions represent the silent mode.

In Fig. 2, we show the ideal situation where an image sequence samples the motion of a line having an arbitrary orientation in the xy-manifold. The line motion will appear

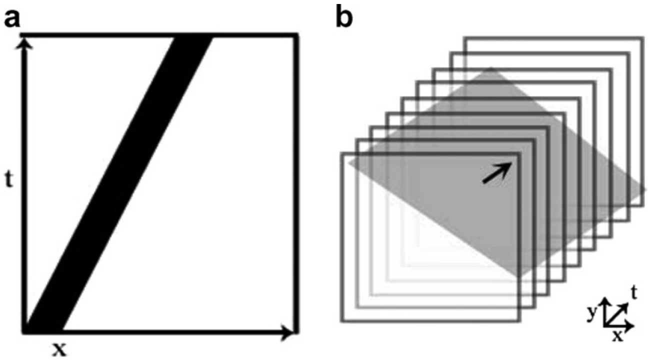


Fig. 2. (a) A line-motion observed in the 2D space–time manifold. (b) A line-motion in the image plane generates the dark plane, in the 3D space time image sequence – xyt, with a normal (marked with black arrow).

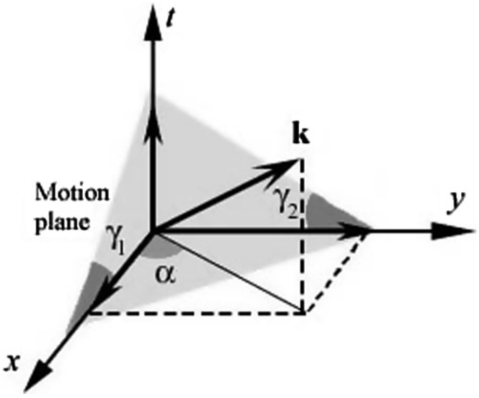


Fig. 3. Illustration of a motion plane in 3D space with $\boldsymbol{k}(\nabla \boldsymbol{f})$ representing the normal of the plane.

as inclined lines in both yt- and xt-planes, Fig. 2a. The line motion generates a (dark) plane in the 3D space–time with a normal vector, Fig. 2b. We describe now the algorithm which estimates the normal velocity of the line-motion from two orientations, in the xt- and yt-manifolds.

Because the motion of a moving line in a spatio-temporal image generates a plane in 3D, it generates a line in the 2D space–time manifolds. The normal velocity of a moving line as it appears to the observer (in the xy-image plane) is in turn determined by the normal of the plane in 3D. Assume that the spatio-temporal plane has a normal $\nabla \boldsymbol{f} = (\mathrm{d}f/\mathrm{d}x, \mathrm{d}f/\mathrm{d}y, \mathrm{d}f/\mathrm{d}t)$, also denoted[1] as $\nabla \boldsymbol{f} = (f_x, f_y, f_t)^{\mathrm{T}}$. The 3D vector $\nabla \boldsymbol{f}$ is orthogonal to the iso-gray surface of $\boldsymbol{f}$ at (x, y, t).

By the shaded area (Fig. 3), we show the motion plane together with its normal $\nabla \boldsymbol{f}$, represented in the figure by its estimation, $\boldsymbol{k}$. The angle α represents the direction of the moving line in the image plane. Assuming that the normal of the tilting plane is $\boldsymbol{k}$, the problem is how the 2D normal flow vector in the xy-manifold can be obtained from the 3D vector. In this case we have a linearly symmetric (local) image in 2D, i.e. in xt- and yt-manifold. Linearly symmetric 2D images are defined as images consisting of iso-gray curves that are (parallel) lines:

$$g(\cos(\alpha)x_0 + \sin(\alpha)y_0) = g(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{s}_0) \tag{1}$$

Here $g(\tau)$ is a one-dimensional function and the vectors

$$\boldsymbol{s}_0 = (x_0, y_0)^{\mathrm{T}}, \quad \boldsymbol{a} = (\cos(\alpha), \sin(\alpha))^{\mathrm{T}} \tag{2}$$

represent the coordinates of an arbitrary point in the image plane and the normal of the line(s) defining the linearly symmetric image, respectively. Eq. (1) manufactures a 2D image by replacing the argument of the 1D function g with the "equation" of a line:

$$\cos(\alpha)x_0 + \sin(\alpha)y_0 = \tau \tag{3}$$

Clearly, the gray-value does not change as long as we are on a line, i.e. the (x_0, y_0) pair satisfies Eq. (3), and therefore the motion of linearly symmetric image is justified when speaking about moving lines in image sequences. We produced a 2D function from a 1D function, by performing a coordinate transformation (CT), because we substituted the single scalar argument of $g(\tau)$ with an expression of two variables. We take the CT one step further and translate one of the lines in the pattern $\cos(\alpha)x_0 + \sin(\alpha)y_0$ using a time parameter t.

We can assume that the position vector $\boldsymbol{s}_0 = (x_0, y_0)^{\mathrm{T}}$ represents a (spatial) point in the image plane at the time instant $t = 0$. We wish to move this line with the velocity $v\boldsymbol{a}$, where $\boldsymbol{a}$ is the direction of the velocity ($\|\boldsymbol{a}\| = 1$) and v is the absolute speed. A velocity in the direction orthogonal to $\boldsymbol{a}$, i.e., when the line moves "along itself", will not be observable which is also known as the *aperture problem*. Accordingly, only in the direction of $\boldsymbol{a}$ can a motion be observed in a linearly symmetric image. It may not be the true motion, but it is the only motion that we can observe. Accordingly, after time t, a point on the line can be assumed to have moved to the position

$$\boldsymbol{s}(t) = (x(t), y(t))^{\mathrm{T}} = \boldsymbol{s}_0 + vt \cdot \boldsymbol{a} \tag{4}$$

so that $\boldsymbol{s}_0 = \boldsymbol{s} - vt \cdot \boldsymbol{a}$. The vector $v\boldsymbol{a}$ is the *normal image velocity* or *normal optical flow*. Substituting the gray-values expression in Eq. (1) yields a spatio-temporal image (sequence) in which the lines of g move with the same velocity:

$$g(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{s}_0) = g(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{s}(t) - vt\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}) = g(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{s} - vt) = g(\tilde{\boldsymbol{k}}^{\mathrm{T}}\boldsymbol{r}) \tag{5}$$

Here we have defined the new variables $\tilde{\boldsymbol{k}}$ and $\boldsymbol{r}$ as the spatial variables augmented with the temporal variables $-v$ and t, respectively $\tilde{\boldsymbol{k}} = [\boldsymbol{a}^{\mathrm{T}}|-v]^{\mathrm{T}} \in E_3$, $\boldsymbol{r} = [\boldsymbol{s}^{\mathrm{T}}|t]^{\mathrm{T}} \in E_3$. However, even Eq. (5) represents a linearly symmetric image but in 3D, i.e. its iso-surfaces consist of parallel planes. The vector $\tilde{\boldsymbol{k}}$ is thus equal to the normal of the plane $\tilde{\boldsymbol{k}}^{\mathrm{T}}\boldsymbol{r} = \text{constant}$ that will be fit by the most significant eigenvector of the 3×3 structure tensor of $f(x, y, t)$ (Bigun and Granlund, 1987). Notice that the first two elements of $\tilde{\boldsymbol{k}}$ are normalized to have length 1. Accordingly, given that at least one of its first two elements is nonnil, the normal vector $\boldsymbol{k}$ is related to $\tilde{\boldsymbol{k}}$ as

$$\tilde{\boldsymbol{k}} = \frac{\boldsymbol{k}}{\sqrt{k_x^2 + k_y^2}} \tag{6}$$

Obtained via such a normalization from the most significant eigenvector of the structure tensor, the first two elements of $\tilde{\boldsymbol{k}}$ will then be equal to $\boldsymbol{a}$:

$$\boldsymbol{a} = \left(\frac{k_x}{\sqrt{k_x^2 + k_y^2}}, \frac{k_y}{\sqrt{k_x^2 + k_y^2}}\right)^{\mathrm{T}} \tag{7}$$

and the third element will be equal to the speed in the normal direction (in the image plane):

[1] Note that the normal velocity in the xy-manifold is invariant to a sign change of the gradient, i.e. the vectors $\nabla \boldsymbol{f}$ and $-\nabla \boldsymbol{f}$ represent the same plane encoding a single velocity vector in the xy-plane.

$$v = \frac{k_t}{\sqrt{k_x^2 + k_y^2}} \tag{8}$$

to the effect that the velocity or the *normal optical flow* will be given by $-v\boldsymbol{a}$

$$\boldsymbol{v} = -v\boldsymbol{a} = -\frac{k_t}{k_x^2 + k_y^2}(k_x, k_y)^{\mathrm{T}}$$
$$= -\frac{1}{\left(\frac{k_x}{k_t}\right)^2 + \left(\frac{k_y}{k_t}\right)^2}\left(\frac{k_x}{k_t}, \frac{k_y}{k_t}\right)^{\mathrm{T}} = (v_x, v_y)^{\mathrm{T}} \tag{9}$$

Assuming the equation of the motion plane is given by $k_x x + k_y y + k_z z = C$, where C is some constant, the intersections of this plane with the x, y, t-axes are given by C/k_x, C/k_y, and C/k_z, respectively, Fig. 3. Accordingly, if we know the tilts of the motion plane in the xt- and yt-manifolds, i.e.

$$\tan\gamma_1 = \frac{k_x}{k_t} \quad \text{and} \quad \tan\gamma_2 = \frac{k_y}{k_t} \tag{10}$$

we can determine the normal velocity, $\boldsymbol{v}$. The tilts $\tan\gamma_1$ and $\tan\gamma_2$ can be estimated in the *total least square* (TLS) error sense as the local directions of the 2D lines in the xy- and yt-manifolds by using the following complex convolution (Bigun et al., 1991; Bigun, 2006).

$$\tilde{u}_1 = \int\int \left(\frac{\partial f}{\partial x} + \mathrm{i}\frac{\partial f}{\partial t}\right)^2 \mathrm{d}x\,\mathrm{d}t \tag{11}$$

$$\tilde{u}_2 = \int\int \left(\frac{\partial f}{\partial y} + \mathrm{i}\frac{\partial f}{\partial t}\right)^2 \mathrm{d}y\,\mathrm{d}t \tag{12}$$

It is worth noting that the quantities $\tilde{u}_1$ and $\tilde{u}_2$ are complex valued and that the "~" denotes that these are TLS estimations of the true directions. Here f is the space–time image that depends on three variables, but the integrations are carried out in 2D manifolds, xt and yt. The obtained complex numbers $\tilde{u}_1$ and $\tilde{u}_2$ correspond to the most significant eigenvectors of the respective 2D structure tensors. They estimate the directions of the lines in the xt- and yt-manifolds, but in the double angle representation (Granlund, 1978). To be precise, the complex numbers $\tilde{u}_1$ and $\tilde{u}_2$ will relate to γ_1 and γ_2 as follows:

$$\tilde{u}_1 = m_1(\cos(2\gamma_1) + \mathrm{i}\sin(2\gamma_1)) = m_1\exp(\mathrm{i}2\gamma_1) \tag{13}$$

$$\tilde{u}_2 = m_2(\cos(2\gamma_2) + \mathrm{i}\sin(2\gamma_2)) = m_2\exp(\mathrm{i}2\gamma_2) \tag{14}$$

where m_1 and m_2 are certainty measures. In consequence, the arguments of $\tilde{u}_1$ and $\tilde{u}_2$ must be halved to yield the two tilt angles, γ_1 and γ_2 providing for an approximation of the velocity, Eqs. (9) and (10).

$$\frac{k_x}{k_t} = \tan\gamma_1 = \tan\left(\frac{1}{2}\arg(\tilde{u}_1)\right) \Rightarrow \tilde{v}_x = \frac{\tan\gamma_1}{\tan^2\gamma_1 + \tan^2\gamma_2} \tag{15}$$

$$\frac{k_y}{k_t} = \tan\gamma_2 = \tan\left(\frac{1}{2}\arg(\tilde{u}_2)\right) \Rightarrow \tilde{v}_y = \frac{\tan\gamma_2}{\tan^2\gamma_1 + \tan^2\gamma_2} \tag{16}$$

Here, the "~" is used to denote that these quantities are estimations of v_x and v_y.

In our implementation we first used Eqs. (11) and (12) to compute the two direction angle components needed to obtain the tilts, Eqs. (15) and (16), which in turn enabled us to estimate the *normal image velocities* in lip-images, Eq. (9). In that, only processing along two planes embedded in 3D spatio-temporal images were needed. In the next section we quantify the accuracy of this motion estimation scheme.

3. Quantification of the motion estimation accuracy

In this section we quantify the velocity estimation algorithm by using two synthetic images, a rotated fan and an expanding circle with different velocities and spatial frequencies. An advantage of synthetic image sequences is that the ground truth regarding velocities is known and an accuracy quantification of velocity vector computations is possible. Here we applied the described algorithm to a rotated fan and expanding circle to compute the motion from images.

The following steps are used for extracting the *normal image velocities* from an image sequence, $f(x, y, t)$.

(I) Permute the space–time image f along the vertical and horizontal axis to obtain the xt-image and yt-image sets.
(II) Calculate $\tilde{u}_1$ and $\tilde{u}_2$ (using Eqs. (11) and (12)) at every pixel of the spatio-temporal image.
(III) Calculate the velocity $\tilde{v}_x$ and $\tilde{v}_y$ from $\tilde{u}_1$ and $\tilde{u}_2$, according to Eqs. (15) and (16).
(IV) Form an image sequence pair to represent the normal velocity, $\boldsymbol{v}(x, y, t) = (v_x(x, y, t), v_y(x, y, t))$.

In the following tests we applied this scheme to quantify the directions and the magnitudes of the *normal image velocity* estimations. All original images have an intensity dynamic range consisting of the integers in the interval $[0, 255]$.

Fig. 4a shows an image containing all possible directions of sine waves with exponentially decreasing frequency in the radial direction of the circles. In the experiments the sine waves were shifted to generate an image sequence with 64 frames. In Fig. 4b we show the profile along a line indicated in (a) where we can observe the varying spatial frequency in the test image as a 1D graph. Fig. 4c illustrates the obtained *normal optical flow* estimation for one frame. The length of the arrows represents the magnitude of velocity and the gray-values in the background image represent the directions of the estimated velocities. We can see that the gray shift is continuous and monotonous. The velocity direction accuracy is given further precision for the white circle in Fig. 4d, where we observe that the estimated velocity direction follows the true velocity direction very closely since the graph is linear. The absolute speeds increase radially in agreement with the ground truth.

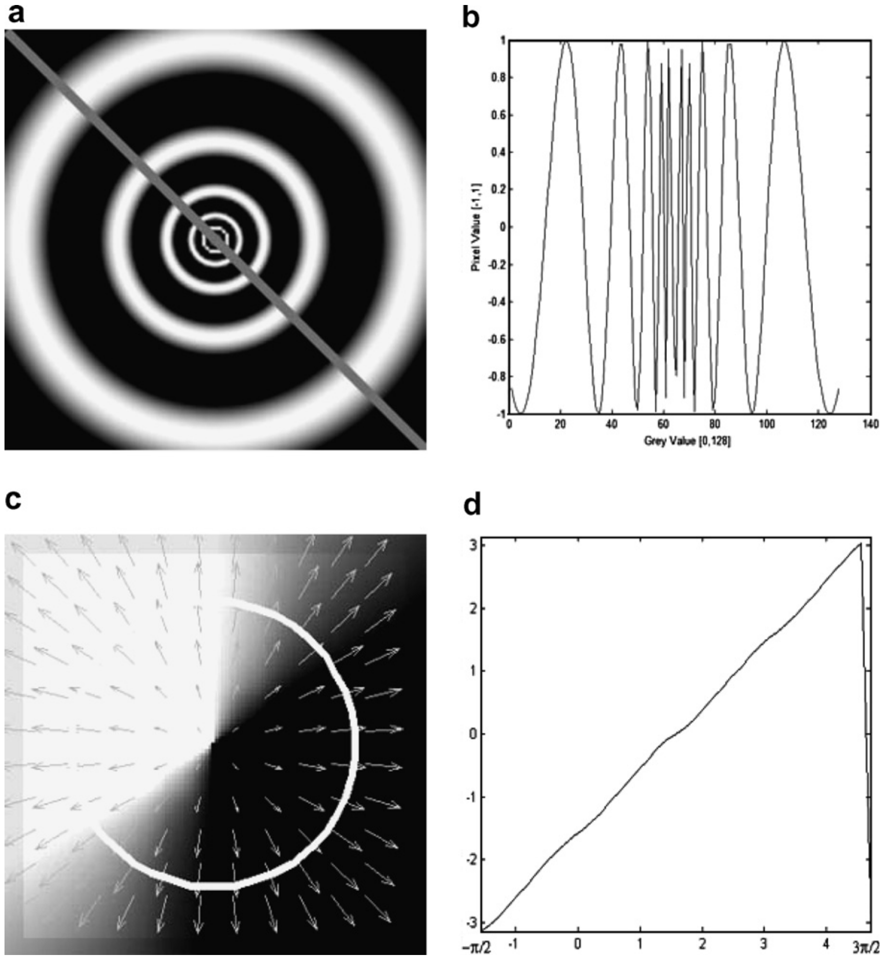


Fig. 4. (a) Expanding waves test image. (b) Profile of (a) along the indicated line. (c) The estimated *normal optical flow* vectors with the velocity direction in background. (d) The graph shows the estimated argument of (c) along the indicated white circle.

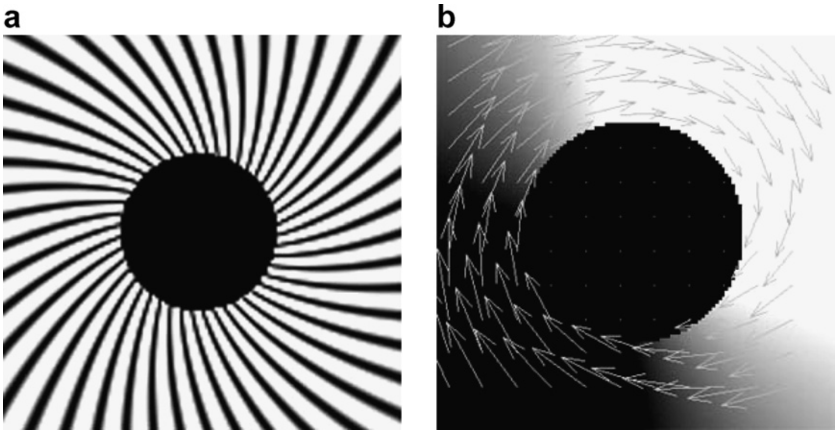


Fig. 5. (a) A skewed fan pattern. (b) The direction and magnitude of *normal image velocity* estimation where the dense background is the estimated velocity direction.

The image in Fig. 5a shows a synthetically produced fan. We rotate the image to create artificial motion and obtain 128 frames. The estimated velocity can be seen in Fig. 5b, where the velocity magnitudes and directions are represented by the gray values and the arrows, respectively. We can see that the arrows change their directions as the directions in a spiral do.

The above results signify a reasonable accuracy of the velocity estimation when local images clearly exhibit line motion and allow us to use such measurements in applica-

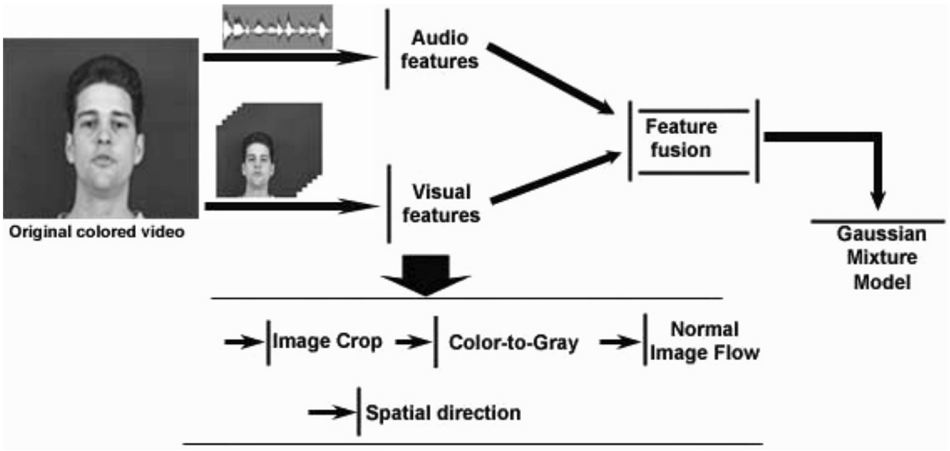


Fig. 6. The suggested joint audio–visual speaker verification system. Speech signal is converted to Mel-Frequency features, which in turn were merged with the *image normal flow* of the lip-motion features. The merged feature sets are then presented to the GMM system for person verification.

tions. In the following section we describe the extracted motion estimation features in a Gaussian mixture model (GMM) framework to perform speaker authentication.

4. Identity verification on XM2VTS

This section describes a speaker verification system using joint modeling of speech and lip-movements in a Gaussian mixture model.

Fig. 6 shows an illustration of our merged system of audio and visual signal for speaker verification. The speech signals were recorded at a 16 kHz sampling rate. The feature extraction is discussed in detail in Section 4.1. The video stream was 720 × 576 in spatial resolution. Before computing the lip-motion the video is manually cropped to yield the lip-area having the size 128 × 128, and the color images are transformed to the gray scale.[2] Section 4.2, describes the visual feature extraction with the steps from Section 2 with further details of the spatial direction. Then next, in Sections 4.3 and 4.4, we present an alternative motion-estimation estimation technique used in our comparisons, and the fusion method, respectively. Finally, in Section 4.5, the speaker verification system setup is presented.

4.1. Speech analysis extracting Mel-frequency features

A person can be distinguished from others by the vocal tract structure, which is implicitly reflected in the speech spectrum. There are several spectral representations of a person's speech. In this paper we use the most common audio features, the cepstral coefficients derived from a Mel-frequency filter-bank. The filter-responses effectively constitute the speech spectrum. From the speech spectrum, cepstral coefficients are extracted, forming our Mel-frequency cepstral coefficients (MFCC) feature vector (Reynolds and Rose, 1995). The sampled speech in the XM2VTS database were stored in files as a common data stream (wave format), which were then processed by using the HTK (Hidden Markov model toolkit) (Young et al., 2000; Veeravalli et al., 2005). The speech features in this study were the MFCC vectors that were generated by the HTK.

The overall process for speech features is illustrated in Fig. 7, which shows the sampled waveform being converted into a sequence of acoustic parameter blocks (A-frame). The duration of the waveform used to compute each parameter vector is usually referred to as a window size, which is set to 25 ms in our experiments. The elapsed time between window size and the output sampling period or frame period is 10 ms. Normally, the window size will be larger than the output sampling period so that successive windows overlap. Each parameter block consists of a 39-dimensional vector. This vector contains 12 cepstral coefficients extracted from the Mel-frequency spectrum of the frame with normalized log energy, 13 delta coefficients, and 13 delta–delta coefficients. The delta and delta–delta coefficients are the first and the second order time derivatives of the extracted cepstral coefficients and are also known as the velocity and the acceleration, respectively.

4.2. Lip-motion features from video

We are interested in person-unique facial changes due to speech production and therefore we analyze the mouth region only. Common approaches to extract such information are often based on geometric features or intensity features, either when the whole face or part of the face are considered (Duc et al., 1997; Sanchez et al., 1997). While it is always present in speech production, it is worth noting that lip-motion can also be present even without speech production, e.g. when producing facial expressions.

[2] Extracting a lip-rectangle, approximately centered on the lip, is possible to do automatically by mouth/eye detection techniques, e.g. see Smeraldi and Bigun (2002). To remove a possible bias of this procedure on the verification significance of lip-movements, we cropped the lip-rectangle manually in the first frame of an image sequence.

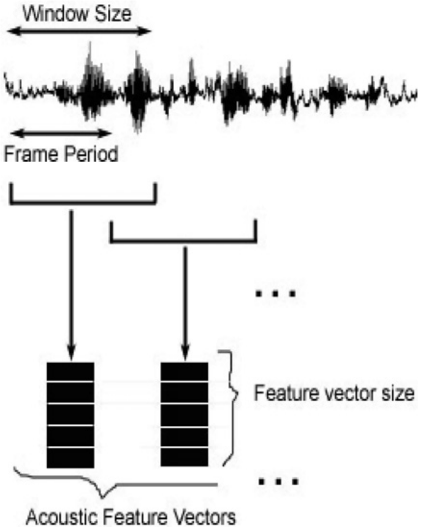


Fig. 7. The feature extraction process with a fixed window size and fixed frame period. Features are extracted for each block forming the speech feature vectors.

Also, lip-reading by motion analysis has been shown to be useful (Fasel and Luettin, 2003; Mase and Pentland, 1991), because during the speech production the lip-deformations and intensity changes in the mouth area change in a highly word-specific manner. In this study we only consider lip-motion due to speech production that is person-specific. We suggest and quantify lip-motion based features around the mouth area in an attempt to obtain person specific lip-dynamics information.

The algorithm described in Section 2 was used to estimate the normal velocities of the lips over each 10 consecutive images of an image sequence for a speaking person where the velocity features were computed in each pixel of the image. Fig. 8 illustrates the *normal optical flow* of the mouth area where we can see that the most significant motion vectors are around the lip-area which also contains most of the edges.

4.3. Motion estimation by differentials in two frames

First, we compare our method, described in Section 2, with *Lucas and Kanade's* approach (Lucas and Kanade, 1981). In this technique, the *optical flow* can be estimated by the solution of a linear regression problem. This happens typically for a local image pattern $g(x,y)$ wherein all points translate with the same velocity. The linear regression problem is also known as *mean square error* (MSE) estimation which results in solving a 2×2 system of equations. The solution involves the inverse of the 2D structure tensor, which uniquely exists if the structure tensor is non-singular, which in turn occurs if and only if the image lacks linear symmetry. A linear symmetry exists in a 2D image if its iso-curves consist of parallel lines. This means that, according to this technique, no velocity will be estimated from the image if an eigenvalue of the structure tensor is zero. Therefore, no velocity will be estimated reliably, if the local image motion is similar to the one illustrated in Fig. 2. This is because the method is designed to estimate the motion of "points", not "lines".

In Fig. 9, we show the estimated *optical flow* (Lucas and Kanade, 1981), on two different speakers from the XM2VTS. Fig. 9a shows a frame of a speaking male (upper figure part) and female (lower figure part) person from the XM2VTS database. By using the two consecutive frames of (Fig. 9b and c), the *optical flow* has been computed and is shown in Fig. 9d.

The quality differences in the motion estimation in the two results are significant because there are far fewer reliably estimated motion vectors for the female person. The *optical flow* quality in the male person images is superior because of the beard-texture, resulting a severe limitation for its use in this application, since the persons without facial hair around their lips are in clear majority (females, and males without facial hair). Also, the technique will not be very useful as significant person-specific information is encoded in the motion of lip-outlines.

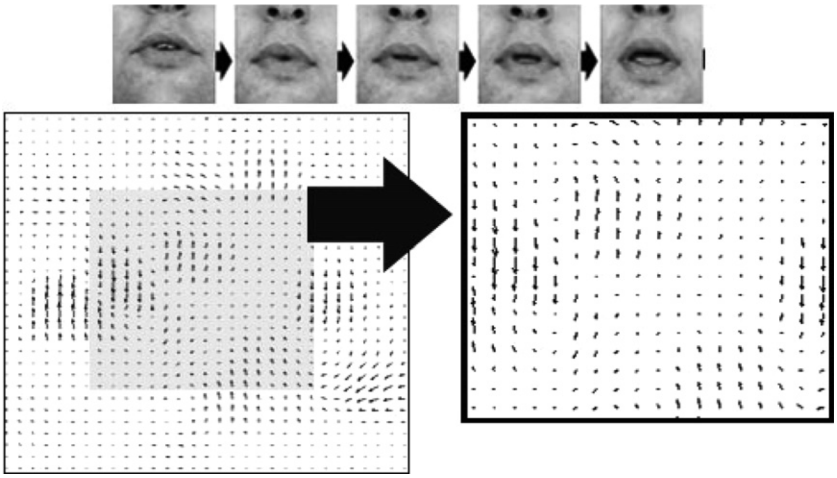


Fig. 8. The estimated *normal image velocity* vectors of lip-movement over some images from an image sequence of a speaker pronouncing "0–9" from the XM2VTS database.

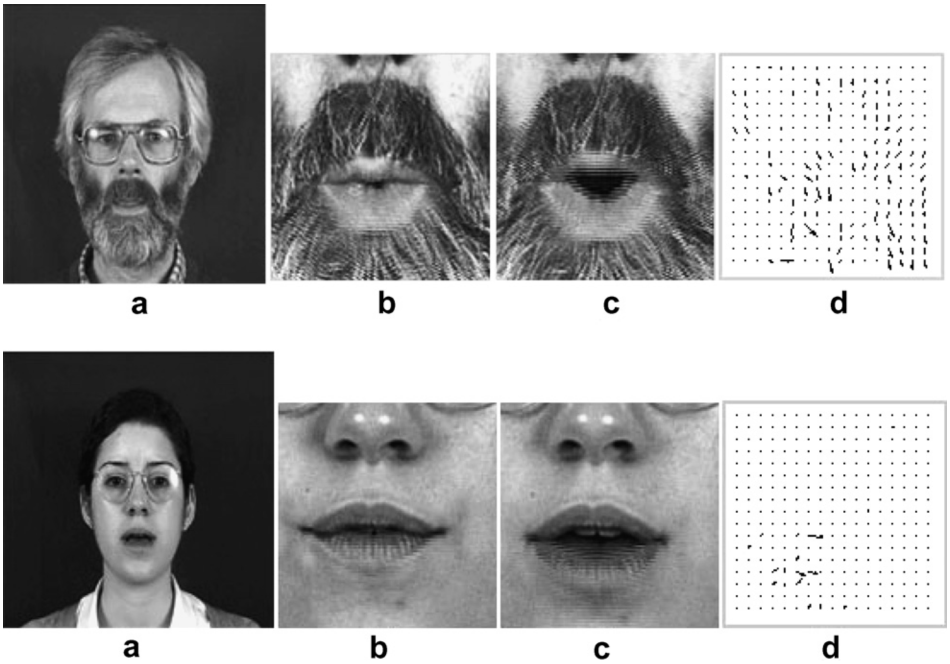


Fig. 9. (a) A male (upper) and female (lower) speaker pronouncing "0–9" from the XM2VTS database. The used mouth area is shown in (b) and (c). (d) The estimated *optical flow* using Lucas and Kanade (1981).

4.3.1. Feature clustering

Using the technique suggested in Section 2, in each mouth-region frame (128 × 128 pixels) we have dense 2D velocity vectors. We need to extract statistical features from the normal velocity to reduce the amount of data without degrading identity-specific information excessively. First, we reduce the 2D vectors to 1D scalars by only allowing three directions (0°, 45°, −45°) as marked with the six solid lines in six regions (Fig. 10a). The motion vectors within each region become real scalars that take the signs + or − depending on which direction they move relative to their expected spatial directions (solid lines). The next step is to quantize the estimated velocities from being allowed arbitrary real scalars to a more limited set of values, here 20. Empirically we found that direction and speed quantization are significant to identity verification as this reduces the impact of noise on the motion information around the lip-area. The quantized speeds are obtained from the data by applying an automatic clustering technique, the fuzzy c-means (Pal and Bezdek, 1995), at four regions of the mouth-region (Fig. 10b). The obtained cluster-centers, and their corresponding cluster-populations, were used as a feature vector for each of the four regions. Consequently, each of these sub-regions has a 40 dimensional feature vector, consisting of 20 cluster-centers and 20 cluster-populations, representing the statistics of lip-motion.

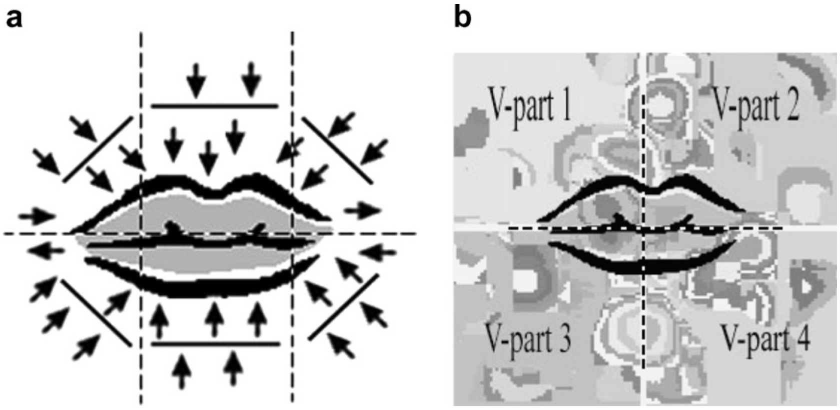


Fig. 10. Illustration of the model used for motion simplification around the mouth area in a lip-movement sequence. (a) The velocity vectors are divided into six regions, marked by dashed lines where each region is projected into a spatial direction marked by the solid line. (b) The results of a clustering of the estimated velocity vectors that were divided onto four parts by the dashed lines. The gray-values encode the absolute speeds in predefined directions.

4.4. Audio–visual fusion

Fusion is a known problem and has been studied increasingly in multi-modal person authentication. It has been investigated in two ways (Varshney, 1997), *feature fusion* and *decision fusion* also called *score fusion*. *Feature fusion* may result in high dimensional feature vectors but it is a more informative way of representing the individual traits than the *decision fusion* because a score or a decision is a real valued scalar whereas a feature is typically represented by a large vector.

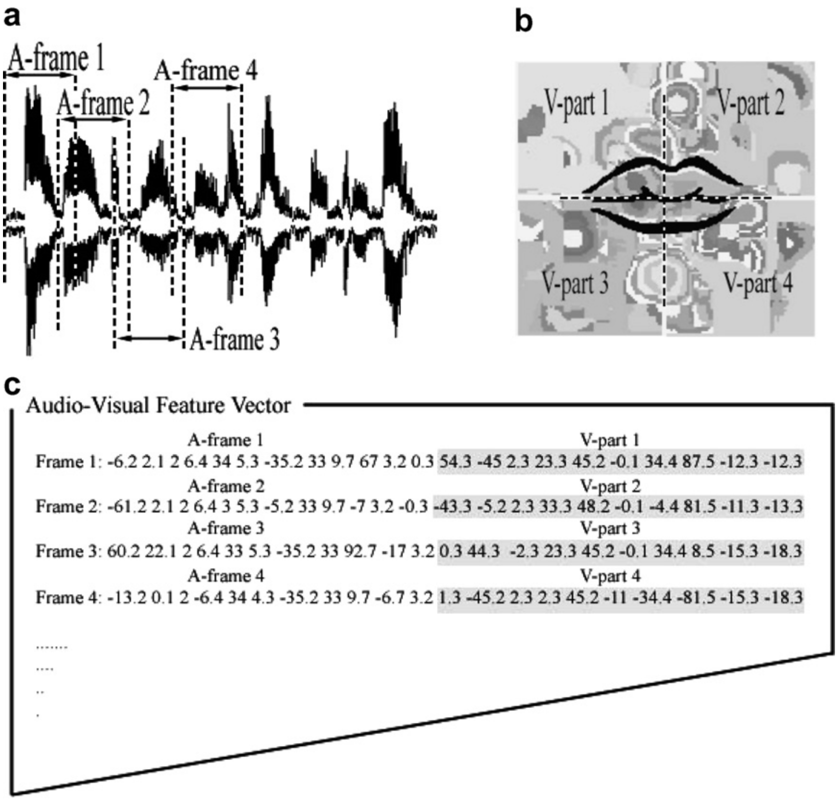
To preserve the discriminatory information as long as possible in the processing chain, we have taken the *feature fusion* approach when combining the audio and video information (Stover et al., 1996). We merge the features that we described above into a single audio–visual feature vector, as illustrated by Fig. 11. This allows us to develop a joint audio–visual dynamic-model for person-specific information in the data. Furthermore, the recognition methods developed for automatic speech and speaker recognition over three decades can be utilized in a straightforward manner. The acoustic and visual sampling rates are, however, different. The speech data as well as visual data are significantly reduced by the feature extraction processing giving an opportunity to synchronize the two data strands at the feature level. The speech feature vectors, described in Section 4.1, covering a window of 25 ms, is shown in Fig 11a. As mentioned, each visual feature vector corresponds to one fourth of a lip-frame (Fig 11b). The factor four is also the ratio between the audio and the visual data output rates to the effect that we can merge each of the four feature vectors of a visual frame with its own audio feature vector (Fig 11c). That is, the merged feature vectors (79 elements) come at the rate of the audio feature vectors but have both audio (39 elements) and lip-motion (40 elements) information. The lip-motion information originating from the same instant are thus distributed in four consecutive samples of the merged data.

The GMM system operates independently of the fusion model. The experiments as to how alternative combinations of the divided frames in the visual signal merged with the audio signal influence the performance will be presented in Section 5.3.

4.5. Speaking person verification using Gaussian mixture model

A Gaussian mixture model can be represented as a weighted sum of multivariate Gaussian distributions (Reynolds and Rose, 1995).

$$p(\boldsymbol{x}|\lambda) = \sum_{j=1}^{M} p_j b_j(\boldsymbol{x}) \tag{17}$$



A-frame 1 V-part 1
Frame 1: -6.2 2.1 2 6.4 34 5.3 -35.2 33 9.7 67 3.2 0.3 54.3 -45 2.3 23.3 45.2 -0.1 34.4 87.5 -12.3 -12.3
A-frame 2 V-part 2
Frame 2: -61.2 2.1 2 6.4 3 5.3 -5.2 33 9.7 -7 3.2 -0.3 -43.3 -5.2 2.3 33.3 48.2 -0.1 -4.4 81.5 -11.3 -13.3
A-frame 3 V-part 3
Frame 3: 60.2 22.1 2 6.4 33 5.3 -35.2 33 92.7 -17 3.2 0.3 44.3 -2.3 23.3 45.2 -0.1 34.4 8.5 -15.3 -18.3
A-frame 4 V-part 4
Frame 4: -13.2 0.1 2 -6.4 34 4.3 -35.2 33 9.7 -6.7 3.2 1.3 -45.2 2.3 2.3 45.2 -11 -34.4 -81.5 -15.3 -18.3

....
....
..
.

Fig. 11. *Direct fusion* of speech and visual features. The speech features (a) are sampled at every 10 ms with a window width of 25 ms. The clustered visual features (b) are sampled at 25 frames/s. The fusion, (c), pairs every speech feature-vector with one visual feature-vector that only represents the lip-motion of one quarter of an image frame.

Here $\boldsymbol{x}$ is a D-dimensional feature vector, p_j and $b_j(\boldsymbol{x})$ for $j = 1 \ldots M$ represent the mixture weights and the component densities that are multivariate Gaussian densities. The Gaussian mixture model is parameterized by the mean vector μ_j, covariance matrix Σ_j and mixture weights p_j for all components $b_j(\boldsymbol{x})$ for $j = 1 \ldots M$. These are the model parameters, and are collectively represented by λ

$$\lambda = \{\mu_j, \Sigma_j, p_j\}, \quad j = 1 \ldots M \tag{18}$$

For each identity a unique GMM is built, i.e. a λ is computed which can also be used to label people's identities. The mixture weight p_j can also be interpreted as the probability that a personal feature $\boldsymbol{x}$ is generated by the component density j. Finally, the function $p(\boldsymbol{x}|\lambda)$ represents the probability that the feature $\boldsymbol{x}$ is observed given that the identity is λ.

The problem of person verification can then be formulated as finding the identity model which has the highest a posteriori probability for an observed sequence of features $\boldsymbol{O} = \{x_1, x_2, \ldots, x_I\}$. Using the Bayes rule, this optimization problem can be reformulated as

$$\arg\max_i \Pr(\lambda_i|\boldsymbol{O}) = \arg\max_i \frac{p(\boldsymbol{O}|\lambda_i)\Pr(\lambda_i)}{p(\boldsymbol{O})} \tag{19}$$

where each speaker is represented by a GMM model λ_i. By assuming, that $p(\boldsymbol{O})$ is the same for all speaker models and that $\Pr(\lambda)$ is equally likely for all person identities we can simplify Eq. (19) to

$$\arg\max_i p(\boldsymbol{O}|\lambda_i) \tag{20}$$

However, the probability $p(\boldsymbol{O}|\lambda)$ can be calculated as the sum of the conditional probabilities $p(\boldsymbol{O}|I,\lambda)$ over all possible states I, and can be written as

$$p(\boldsymbol{O}|\lambda) = \sum_I p(\boldsymbol{O}, I|\lambda) = \sum_I p(\boldsymbol{O}|I,\lambda)p(I|\lambda) \tag{21}$$

The recognition problem is then equivalent to maximizing equation (21) and involves the following three problems and their well studied solutions (Furui, 1997):

(I) *Evaluation*: How to compute $p(O|\lambda)$ given λ? This is achieved by the forward–backward algorithm.
(II) *Decoding*: How to choose I so that $p(O|I,\lambda)$ is maximized? This is achieved by the Viterbi algorithm.
(III) *Estimation*: How to estimate the model λ? This is archived by the Baum–Welch algorithm.

The training process, described in Section 4.6, was carried out by the HTK using the Baum–Welch re-estimation technique and this improves the recognition results on training set incrementally. The use of this method can be described as an iterative refinement (in the maximum likelihood sense) meaning that the parameters of the model are changed stepwise (Young et al., 2000).

4.5.1. Normalization

When recording dynamic signals, in particular speech, for analysis, the problem of variation of the signal arises for different samples over the time. This variability over time or even within the same session can be due to the source itself (speaker), the way of recording, or transmission noise. Therefore a normalization of the results is needed. In (Furui, 1997), a commonly used technique is described to achieve invariance or resilience against the above variations, which we also used.

- First: Average the MFCC over one utterance and subtract these values from the coefficients of each block.
- Second: Normalize the distance or likelihood estimation, e.g. normalization for the likelihood ratio which is calculated between the conditional probability given that the identity claim is authentic (client) and the conditional probability given that the speaker is an impostor (false claim).

4.6. System setup (HTK)

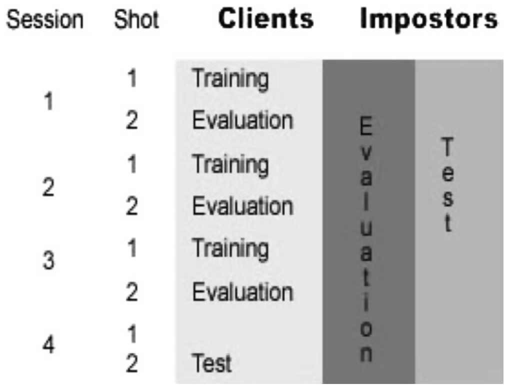
The various steps involved in all our recognition systems are carried out by the following steps:

- Feature preparation
  - Partition the database for training, evaluation, and testing
  - Speech parameter extraction
  - Visual parameter extraction
- Define GMM's structure
- Train the GMM models
  - Single left to right state constellation using a Gaussian mixture model with five states and three mixtures in each state
  - Flat start, which means an unsupervised learning
  - Training process using Baum–Welch re-estimation
  - Forward–backward algorithm (re-estimation of the training process)
- Recognition/performance evaluation
  - Test the current data against reference data set by the Viterbi algorithm
- Adapt the dataset
  - Adaptation is used to reduce the mismatch between the current data and the model set due to source, sensor, and environment variations. A maximum likelihood linear regression (MLLR) is applied to adapt a model set to the current data.

These steps are implemented in HTK software environment (Young et al., 2000; Veeravalli et al., 2005).

5. Experimental test and evaluation of speaker verification

Here, a *text-dependent* GMM speaker verification system is suggested, for its additional usefulness in



Session	Shot	Clients	Impostors	
1	1	Training	Evaluation	Test
	2	Evaluation		
2	1	Training		
	2	Evaluation		
3	1	Training		
	2	Evaluation		
4	1	Test		
	2			

Fig. 12. Partitioning of the XM2VTSDB database according to the Lausanne protocol – Configuration I.

verifying "liveness"[3] as a measure against play-back or other spoofing attacks. The system was evaluated in a task where the speech features and the visual features were combined as in Section 4.3. The XM2VTS database along with experimental results on this database are presented first.

5.1. Database description and experimental protocol

The XM2VTS audio–visual database (Messer et al., 1999), contains audio and video sequences for 295 speakers (male and female). For each person, several video sequences are taken over four different sessions. In each session, a person is asked to pronounce "0 1 2 3 4 5 6 7 8 9", "5 0 6 9 2 8 1 3 7 4" when recording the video sequence. The Lausanne protocol (Luettin and Maitre, 1998) or the XM2VTS protocol is a common experimental procedure for speaker verification and identification using different modalities. The database is divided, according to Configuration I (Fig. 12) of the Lausanne protocol, into training, evaluation and test sets.

In Fig. 12, the training set contains 200 speakers and 70 speakers as impostors for test and 25 speakers as impostors for evaluation. The first recoding of each sentence of session 1–3 is used as training set and the second recording of session 1–3 is the evaluation set. Session 4 is used as a test set. As impostors, 25 speakers were used for evaluation and 70 speakers for testing. The evaluation set is used to produce client and impostor access scores which were used to find the threshold for accepting or rejecting a person.

5.2. Performance quantification

To investigate the person verification performance of the measured features, the following experiments were conducted on all speakers. The false acceptance rate FA and the false rejection rate FR were calculated as follows:

$$\mathrm{FA} = (EI/I) * 100 \tag{22}$$

$$\mathrm{FR} = (EC/C) * 100 \tag{23}$$

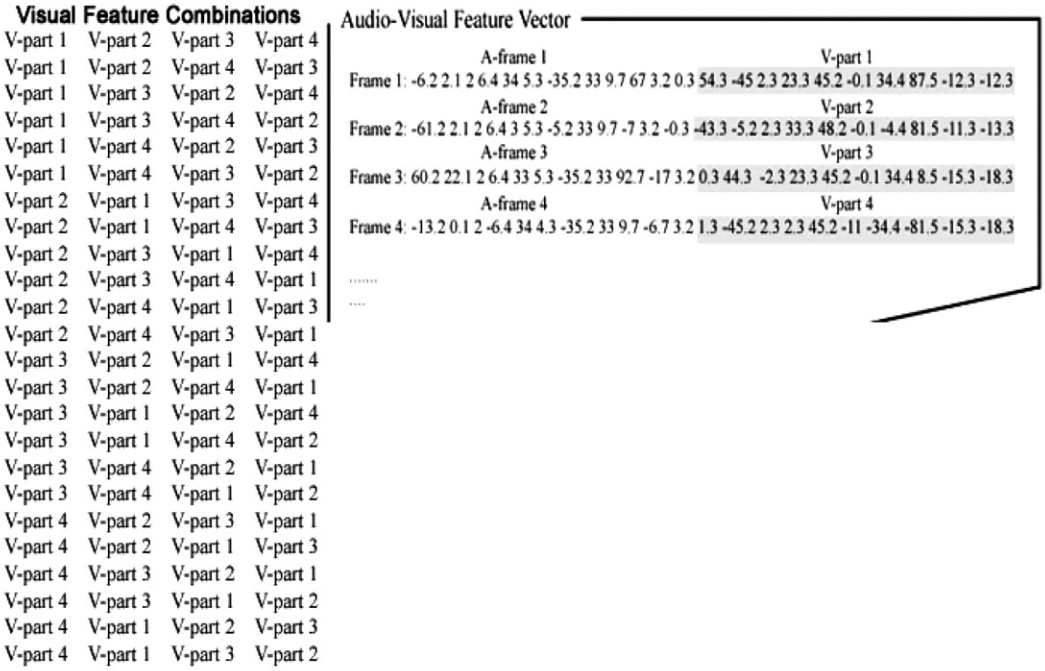
Here the number of impostor and impostor acceptance are denoted with I and EI, and the number of clients and client rejections are denoted with C and EC. Eqs. (22) and (23) are computed on the evaluation set to compute the threshold for which the number of false acceptance and false rejection errors are equal. This threshold of the *equal error rate* (EER) is then used on the test set. The FA is marked by the dotted line and the FR is marked by the dashed line in the experimental graphs, detailed in Section 5.4.

5.3. Experimental evaluation of the synchronization/association method

This experiment is carried out to evaluate the association technique which also attempts to solve the synchronization problem by spatio-temporal association maps of the image and the sound data. An equally important issue in this effort was whether or not the various combinations exhibit significantly different verification performance. This is because identity specific visual motion is likely to be present in all four subparts of a lip-image and that these features are highly correlated. We evaluated the latter hypothesis by performing *feature fusion* of audio frames with all possible visual frames based on a GMM audio–visual system. First we used 50 speakers and later the entire database, as explained in Section 4, in this effort.

A verification test was thus carried out on all 24 possible combinations of the visual features (Fig. 13(left)) to be fused with the audio features. The second combination of this audio–visual *feature fusion*, Fig. 13(right), gave the highest EER and the third combination the lowest EER when using 50 speakers, though with insignificant difference. In Table 1, we see the audio–visual verification results of the various combinations where combination 3 gave the lowest EER and combination 2 gave the highest EER using the whole database (295 subjects). Again, the difference between the best and worst case of combination is small even when using the entire database. This confirms the hypothesis that because the motion in a lip-image sequence is roughly symmetrical, the particular *feature fusion* combination has an insignificant effect at the person verification level. Possibly a true asymmetry in the lip-motions could cause stronger verification features for certain individuals, if combinations to be fused were adapted to be person-specific. However, to design and test such a system would require much more data and processing than is currently practicable. Because of the small difference to the average, we conclude that the specific association of the visual features with audio is a less important issue in performance evaluation of our system, and therefore we report on and discuss here only the first combination (Fig. 13).

[3] A biometric system can increase its barriers against spoofing attacks, by seeking evidence for "liveness" i.e. attempting to detect if the biometric signal is captured from a physically present person.



Visual Feature Combinations

V-part 1	V-part 2	V-part 3	V-part 4
V-part 1	V-part 2	V-part 4	V-part 3
V-part 1	V-part 3	V-part 2	V-part 4
V-part 1	V-part 3	V-part 4	V-part 2
V-part 1	V-part 4	V-part 2	V-part 3
V-part 1	V-part 4	V-part 3	V-part 2
V-part 2	V-part 1	V-part 3	V-part 4
V-part 2	V-part 1	V-part 4	V-part 3
V-part 2	V-part 3	V-part 1	V-part 4
V-part 2	V-part 3	V-part 4	V-part 1
V-part 2	V-part 4	V-part 1	V-part 3
V-part 2	V-part 4	V-part 3	V-part 1
V-part 3	V-part 2	V-part 1	V-part 4
V-part 3	V-part 2	V-part 4	V-part 1
V-part 3	V-part 1	V-part 2	V-part 4
V-part 3	V-part 1	V-part 4	V-part 2
V-part 3	V-part 4	V-part 2	V-part 1
V-part 3	V-part 4	V-part 1	V-part 2
V-part 4	V-part 2	V-part 3	V-part 1
V-part 4	V-part 2	V-part 1	V-part 3
V-part 4	V-part 3	V-part 2	V-part 1
V-part 4	V-part 3	V-part 1	V-part 2
V-part 4	V-part 1	V-part 2	V-part 3
V-part 4	V-part 1	V-part 3	V-part 2

Audio-Visual Feature Vector

A-frame 1 V-part 1
Frame 1: -6.2 2.1 2 6.4 34 5.3 -35.2 33 9.7 67 3.2 0.3 54.3 -45 2.3 23.3 45.2 -0.1 34.4 87.5 -12.3 -12.3
A-frame 2 V-part 2
Frame 2: -61.2 2.1 2 6.4 3 5.3 -5.2 33 9.7 -7 3.2 -0.3 -43.3 -5.2 2.3 33.3 48.2 -0.1 -4.4 81.5 -11.3 -13.3
A-frame 3 V-part 3
Frame 3: 60.2 22.1 2 6.4 33 5.3 -35.2 33 92.7 -17 3.2 0.3 44.3 -2.3 23.3 45.2 -0.1 34.4 8.5 -15.3 -18.3
A-frame 4 V-part 4
Frame 4: -13.2 0.1 2 -6.4 34 4.3 -35.2 33 9.7 -6.7 3.2 1.3 -45.2 2.3 2.3 45.2 -11 -34.4 -81.5 -15.3 -18.3
....
....

Fig. 13. Audio–visual feature combination and fusion. In the left part of the image, all possible combinations of visual feature frames are presented. Right of the image we see the merging method of audio frames with first alternative of the visual features.

Table 1
Verification results of first, second and third possible alternatives of the audio–visual *feature fusion* using 295 subjects

Audio–visual system	Evaluation (%)	Test (%)
Combination 1	99.5	97.8
Combination 2	99.4	97.6
Combination 3	99.6	98.1

5.4. Experimental results

First we present the results for a GMM based person using only acoustic features. The threshold function, obtained from the evaluation set (Fig. 14a), is used in the test set to map the score into the confidence interval [0, 1]. From the verification results on the test set (Fig. 14b), we obtain the minimum *equal error rate* of 6%. The verification rate of 295 speakers was thus 94%. In the second test, the system used the merged visual and acoustic features on the same basis as the earlier system. The verification rate of the whole database was then 98% (Fig. 14d), where the threshold from the evaluation set was utilized (Fig. 14c).

To quantify the biometric verification power of the visual features alone, we carried out the same experiments with these features alone. The FA and FR graphs of the verification results are given in Fig. 15a. The verification performance using the threshold of the evaluation set was 78%. In Table 2, we display the verification performance (where FA is equal to FR in the evaluation set) of the acoustic, visual, and the combined audio–visual systems using the same test data and test protocol. Speaker verification based on audio and visual images from lip-movement gives 98% correct verification which is 4–5% points better than audio and 22% points better than visual speaker verification, Fig. 15b. It is worth noting that replacing our motion estimation technique with that of Lucas and Kanade (1981) yields a combined verification rate of 95%, Fig. 15b, which is 3–4% points worse than using our motion estimator.

The improvement of the combined system compared to the speech only system is a ≈50% reduction in EER. This confirms the importance of the visual signal as a complementary information to speech, not the least because it is more difficult to reduce the verification errors when the speech system has already a low error rate as compared to when it would have a high error rate. The good performance of the speech system is explained by the high quality of the speech data (office environment) in the XM2VTS database. As a consequence, these experiments support the view that the added value of lip-motion, measured as EER reduction of a person recognition system in environments having heavy (acoustic) noise (e.g. airports, trains, airplanes) is likely to be higher than ≈50%, not less. A speaker recognition system deployed in a noisy environment e.g. an airport could therefore potentially benefit a reduction of its errors by an order of magnitude if lip-motion information is added.

5.5. Comparative discussion

We quantified the significance of lip-movements in biometric person authentication as a stand-alone modality as

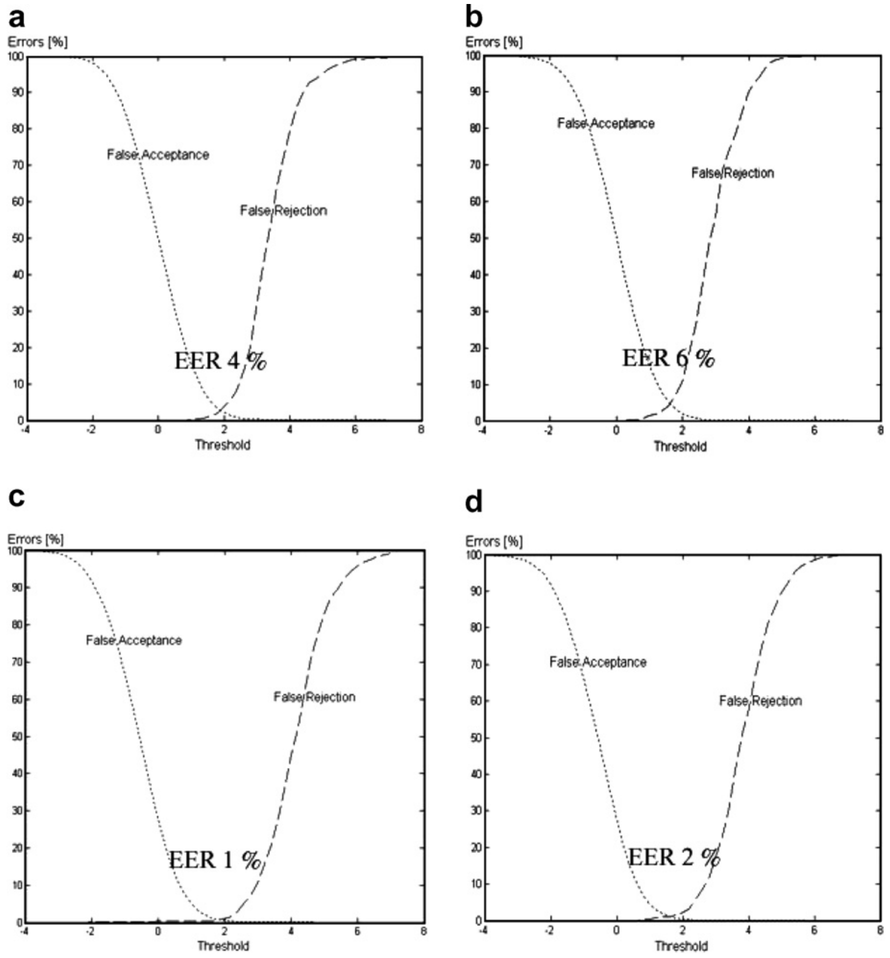


Fig. 14. Speaker verification results using only audio signal and audio–visual signal. Acoustic evaluation results are in (a) and verifications (test) results are in (b). The graphs in (c) and (d) show the corresponding evaluation and test results for the combined audio–visual verification system.

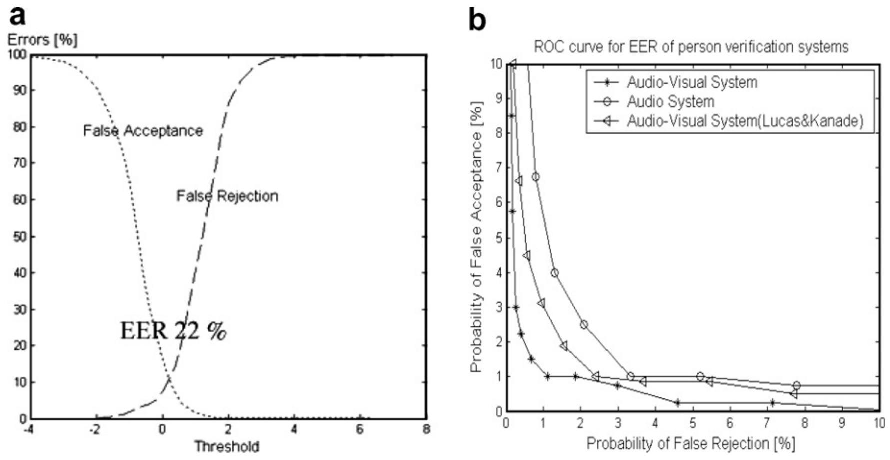


Fig. 15. (a) Verification results of the visual verification system. (b) The ROC curves of different audio and/or visual verification systems.

well as in conjunction with the audio modality using a large (currently the largest) database. The closest comparable study is the system reported by Jourlin et al. (1997). The experiments in that report, however, were carried out on the M2VTS database, containing 37 speakers. They reported 72% verification rate assuming that the tracking of lip-contours was successful in all frames of all image sequences. We report 78% verification rate on lip-motion only, using a significantly larger dataset of 295 persons. Additionally, our technique does not presuppose a successful tracking since the visual features we suggest require neither segmentation nor lip-tracking. Another difference

Table 2
Verification results of the acoustic, visual, and the combined audio–visual systems

Set/system	Evaluation (%)	Test (%)
Acoustic	96	94
Visual	80	78
Audio–visual	99	98

between the two systems is that in the present study we have fused image and audio features, whereas in the above study it was the scores of the image and audio that were fused.

We can only compare our algorithmic elements with other reports that studied the *optical flow* for lip-motion (Dieckmann et al., 1997; Frischholz and Dieckmann, 2000; Tamura et al., 2004), because these studies have not reported verification results on databases that allow a comparison. First, the motion estimation technique used in (Dieckmann et al., 1997; Frischholz and Dieckmann, 2000; Tamura et al., 2004) measures the motion of points as it avoids regions having the aperture problem. Accordingly, this type of *optical flow* presupposes the availability of texture in local images (lack of lines) which is sometimes naturally available, e.g. a beard close to mouth, or unshaved male face, but mostly not. By contrast, the motion estimation technique used here, assumes the opposite, namely, it requires the presence of spatial direction, such as lines and edges (not dots or isotropic texture), in still image frames. Our experiments carried over both female and male persons' data support that these image structures are available in the moving lips in significant amounts in practice. This becomes more important when considering that the XM2VTS database images show entire faces without a particular focus on the mouth region, causing the lip-areas to have relatively low resolution (128×128). Second, our lip-motion features do not contain any iterative algorithm, as all computations are based on closed form arithmetic function evaluations. In commonly used iterative algorithms, similar to Horn–Shunk (Horn and Schunck, 1981), the actual computation time per image frame is not constant. This makes the implementation of iterative methods more difficult on simple computational architectures.

6. Conclusion and discussion

We have presented a motion estimation technique with application to biometric identification systems. The technique exploits information from a set of orientations of projected 2D images from the 3D time-image that yield the normal of an optimal plane which estimates velocities. Our results indicate that the technique's accuracy and reliability to extract discriminative velocities features are sufficient as stand alone and in combination with audio information in biometric identity verification systems. This solution requires neither segmentation nor lip-tracking, which is a computationally efficient alternative to the general lip-dynamics estimations. Furthermore, this technique is the first to suggest lip-motion features for person authentication to the best of our knowledge.

The performance of the suggested biometric system, yielding approximately 80% verification rates for the lip-motion alone, supports the conclusion that our features of lip-dynamic contains significant information for person authentication. Feature integration of extracted motion statistics with audio information improved the speaker verification performance of even high quality acoustic information (office environment) with additional 4–5% points on top of the already high verification rates (≈94%), i.e. reducing the equal error rate with ≈50%. We can with reasonable confidence conclude that the suggested lip-motion features offer relevant measures for person verification purposes. Considering the size of the test database, the results also support the view that the lip-motion information offers a way to improve identity authentication systems using speech in the presence of (acoustic) noise.

We have provided experimental support for an idea of feature integration at an early stage. It indicates better decision performance than score fusion in audio–visual person authentication, while it offers a possibility to verify "liveness" of the users in applications.

In this work we focused on the feature extraction technique with an application for speaker verification system. However, there are still many interesting problems left to investigate. These include the effect of feature reduction method, Gaussian Mixture Model, and the training method. The experimental results of an integrated speech (limited vocabulary) and speaker recognition using lip-motion and speech in a discriminative classifier setting (SVM) is envisaged.

Acknowledgement

We gratefully acknowledge the support of The Swedish National Research Foundation.

References

Bigun, J., 2006. Vision with Direction. Springer, Heidelberg.

Bigun, J., Granlund, G., 1987. Optimal orientation detection of linear symmetry. First International Conference on Computer Vision, ICCV, London, June 8–11. IEEE Computer Society, pp. 433–438.

Bigun, J., Granlund, G., Wiklund, J., 1991. Multidimensional orientation estimation with applications to texture analysis of optical flow. IEEE Trans. Pattern Anal. Machine Intell. 13 (8), 775–790.

Bigun, E., Bigun, J., Duc, B., Fischer, S., 1997. Expert conciliation for multi modal person authentication systems by bayesian statistics. In: Bigun, J., Chollet, G., Borgefors, G. (Eds.), Audio and Video Based Person Authentication – AVBPA97, pp. 291–300.

Brunelli, K.R., Falavigna, D., 1995. Person identification using multiple cues. IEEE Trans. Pattern Anal. Machine Intell. 17 (10), 955–966.

Chen, T., 2001. Audiovisual speech processing. IEEE Signal Process. Mag. 18 (1), 9–21.

Chibelushi, C., Deravi, F., Mason, J., 2002. A review of speech-based bimodal recognition. IEEE Trans. Multimedia 4 (1), 23–37.

Dieckmann, U., Plankensteiner, P., Wagner, T., 1997. Acoustic-labial speaker verification. In: Proc. First Internat. Conf. on Audio- and Video-Based Biometric Person Authentication, LNCS 1206, pp. 301–310.

Duc, B., Fischer, S., Bigun, J., 1997. Face authentication with sparse grid gabor information. IEEE Int. Conf. Acoust. Speech Signal Process. 4 (21), 3053–3056.

Dupont, S., Luettin, J., 2000. Audio–visual speech modelling for continuous speech recognition. IEEE Trans. Multimedia 2 (3), 141–151.

Faraj, M.I., Bigun, J., 2006. Person verification by lip-motion. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshop – CVPR 2006, pp. 37–45.

Fasel, B., Luettin, J., 2003. Automatic facial expression analysis: A survey. J. Pattern Recognition Soc. 36 (1), 259–275.

Frischholz, R., Dieckmann, U., 2000. Bioid: A multimodal biometric identification system. IEEE-Computer Society Press, vol. 33(2), 2000, pp. 64–68.

Furui, S., 1997. Recent advances in speaker recognition. In: Proc. First Internat. Conf. on Audio- and Video-Based Biometric Person Authentication, LNCS 1206, pp. 237–252.

Granlund, G.H., 1978. In search of a general picture processing operator. Computer Graphics Image Process. 8 (2), 155–173.

Hazen, T.J., 2006. Visual model structures and synchrony constraints for audio–visual speech recognition. IEEE Trans. Audio Speech Lang. Process. 14 (3), 1082–1089.

Horn, B., Schunck, B., 1981. Determining optical flow. J. Art. Intell. 17 (1), 185–203.

Jourlin, P., Luettin, J., Genoud, D., Wassner, H., 1997. Acoustic-labial speaker verification. In: Proc. First Internat. Conf. on Audio- and Video-Based Biometric Person Authentication, LNCS 1206, pp. 319–326.

Kollreider, K., Fronthaler, H., Bigun, J., 2005. Evaluating liveness by face images and the structure tensor. In: AutoID 2005: Fourth Workshop on Automatic Identification Advanced Technologies. IEEE Computer Society, pp. 75–80.

Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. Int. Joint Conf. Art. Intell., 674–679.

Lucey, S., Chen, T., Sridharan, S., Chandran, V., 2005. Integration strategies for audiovisual speech processing: Applied to text-dependent speaker recognition. IEEE Trans. Multimedia 7 (3), 495–506.

Luettin, J., Maitre, G., 1998. Evaluation protocol for the extended m2vts database *xm2vtsdb*. In: IDIAP Communication 98-054, Technical report R R-21, number = IDIAP – 1998.

Luettin, J., Thacker, N., Beet, S., 1996. Speaker identification by lipreading. In: Proc. 4th Internat. Conf. on Spoken Language Processing ICSLP'96, pp. 62–65.

Mase, K., Pentland, A., 1991. Automatic lip-reading by optical-flow analysis. Systems Comput. Jpn. 22 (6), 67–76.

Messer, K., Matas, J., Kittler, J., Luettin, J., 1999. Xm2vtsdb: The extended m2vts database. In: Second International Conference of Audio and Video-Based Biometric Person Authentication ICSLP'96, pp. 72–77.

Pal, N., Bezdek, J., 1995. On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Systems 3 (3), 370–379.

Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A., 2003. Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE 91 (9), 1306–1326.

Reynolds, D., Rose, R., 1995. Robust text-independent speaker identification using gaussian mixture models. IEEE Trans. Speech Audio Process. ICASSP090 3 (1), 72–83.

Sanchez, M.R., Matas, J., Kittler, J., 1997. Statistical chromaticity-based lip tracking with b-splines. In: Proc. First Internat. Conf. on Audio- and Video-Based Biometric Person Authentication, LNCS 1206, pp. 69–76.

Smeraldi, F., Bigun, J., 2002. Retinal vision applied to facial features detection and face authentication. Pattern Recognition Lett. 23, 463–475.

Stover, J., Hall, D., Gibson, R., 1996. A fuzzy-logic architecture for autonomous multisensor data fusion. IEEE Trans. Industr. Electron. 43 (3), 403–410.

Tamura, S., Iwano, K., Furui, S., 2004. Multi-modal speech recognition using optical flow analysis for lip images. J. VLSI Signal Process. 36 (2), 117–124.

Tang, X., Li, X., 2001. Fusion of audio–visual information integrated speech processing. In: Third Internat. Conf. on Audio- and Video-Based Biometric Person Authentication AV BPA02001, LNCS 2091, pp. 127–143.

Tang, X., Li, X., 2004. Video based face recognition using multiple classifiers. In: Sixth IEEE Internat. Conf. on Automatic Face and Gesture Recognition FGR2004 – IEEE Computer Society, pp. 345–349.

Varshney, P., 1997. Multisensor data fusion. Electron. Commun. Eng. J. 9 (6), 245–253.

Veeravalli, A.G., Pan, W., Adhami, R., Cox, P.G., 2005. A tutorial on using hidden markov models for phoneme recognition. In: Proc. Thirty-Seventh Southeastern Symp. on System Theory, SSST 2005.

Wark, T., Sridharan, S., Chandran, V., 1999. The use of speech and lip modalities for robust speaker verification under adverse conditions. IEEE Int. Conf. Multimedia Comput. Systems 1.

Yamamoto, E., Nakamura, S., Shikano, K., 1998. Lip movement synthesis from speech based on hidden markov models. J. Speech Commun. 26 (1), 105–115.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2000. The htk book (for htk version 3.0) http://htk.eng.cam.ac.uk/docs/docs.shtml.