(C) IEEE Computer Society, 2006
In Proc. Int. Conf. on Pattern Recognition, ICPR, pp.1059-1062, Hong Kong, Aug 20-24, 2006

# Motion Features from Lip Movement for Person Authentication

Maycel Isaac Faraj and Josef Bigun
School of Information Science, Computer and Electrical Engineering (IDE)
Halmstad University, Box 823, SE-301 18 Halmstad
{maycel.faraj, josef.bigun}@ide.hh.se

## Abstract

*This paper describes a new motion based feature extraction technique for speaker identification using orientation estimation in 2D manifolds. The motion is estimated by computing the components of the structure tensor from which normal flows are extracted. By projecting the 3D spatiotemporal data to 2-D planes we obtain projection coefficients which we use to evaluate the 3-D orientations of brightness patterns in TV like image sequences. This corresponds to the solutions of simple matrix eigenvalue problems in 2D, affording increased computational efficiency. An implementation based on joint lip movements and speech is presented along with experiments which confirm the theory, exhibiting a recognition rate of 98% on the publicly available XM2VTS database.*

## 1. Introduction

Image sequence based speaker recognition systems have recently attracted the research attention. The performance of multimodal systems using audio and visual information is known to be superior to those of the acoustic and visual subsystems [3]. Specifically, recognition systems using visual information from lip movements provide supplementary information, which can lead to improved speaker recognition performance as demonstrated by [5]. The image based system requires, however, robust lip contours extraction which is affected by noise, requiring frequent manual intervention. Another disadvantage is the non constant computation time due to the iterative convergence process of the contours.

This paper describes an algorithm that takes advantage of the low-level spatiotemporal information contained in an image sequence containing lip-motion. Structure in spatiotemporal images is modeled by moving line patterns in space-time planes, where the normal of the plane encodes to the normal velocity of lines. In the next section we will present a method for normal velocity estimation based on 3D spatiotemporal space [2] but only by use of 2D signal processing. In *section* 3, we show usefulness of these by reporting results from, to the best of our knowledge, largest experimental study using joint lip-motion and speech features. The full XM2VTS database [7] containing audio and video has been used for performance evaluation, yielding a recognition rate of 98%.

## 2. Velocity estimation

Motion estimation, also known as optical flow, can be determined by eigenvalue analysis of the structure tensor [2]. This method requires multiple time frames since it simultaneously derives the velocities of points and lines. Assuming that only line motion can be observed in local images, as our experiments indicate lip-motion, the original orientation detection in 3D can be approximated by a combination of orientation detections in 2D planes.

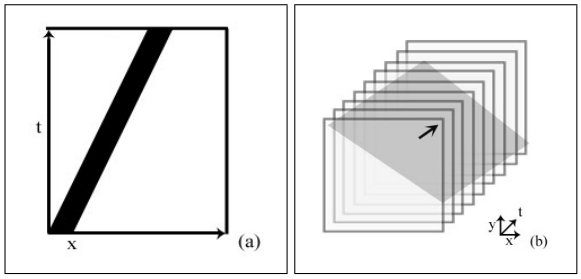


**Figure 1. a) A line-motion observed in the 2D space-time manifold. b) A line-motion in the image plane generates the dark plane in the 3D space time image sequence, $xyt$.**

In Fig. 1, we show the ideal situation where an image sequence samples the motion of a line having an arbitrary orientation in the $xy$-plane. The line motion will appear as inclined lines in both $yt$- and $xt$-planes (Fig. 1a). The line motion generates a grey plane in the 3D space-time mani-

fold, $xyt$, (Fig. 1b). The normal vector of the motion plane is drawn in the figure. We describe now the algorithm which determines the normal velocity of the line-motion from two orientation estimations, in the $xt$- and $yt$- manifolds.

### 2.1. Velocity estimation by 2D orientation detection

The motion of a moving line in a spatiotemporal image generates a plane in 3D but is still a line in the 2D space-time image. The normal velocity in the image plane is determined by the normal of the plane. Assume that the spatiotemporal plane has a normal $\nabla \mathbf{f} = (df/dx, df/dy, df/dt)$, also denoted[1] as $\nabla \mathbf{f} = (f_x, f_y, f_t)^T$. The 3D vector $\nabla \mathbf{f}$ is orthogonal to the iso-gray surface of $\mathbf{f}$ at $(x, y, t)$.

Assume that the normal of the tilting plane is $\mathbf{k} = (k_x, k_y, k_z)$ where the coordinates $x$ and $y$ represent any arbitrary point in the image plane. By a line movement we have a velocity $v\mathbf{a}$ of a moving line in $xyt$-space, where $\mathbf{a}$ is the direction of the velocity ($\|\mathbf{a}\| = 1$)

$$\mathbf{a} = (\frac{k_x}{\sqrt{k_x^2 + k_y^2}}, \frac{k_y}{\sqrt{k_x^2 + k_y^2}})^T \tag{1}$$

and $v$ is the absolute speed in the normal direction (in the image plane):

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

to the effect that the velocity or the *normal optical flow* will be given by $v\mathbf{a}$
$\mathbf{v} = -v\mathbf{a} = -\frac{k_t}{k_x^2+k_y^2}(k_x, k_y)^T =$

$$-\frac{1}{(\frac{k_x}{k_t})^2 + (\frac{k_y}{k_t})^2}(\frac{k_x}{k_t}, \frac{k_y}{k_t})^T \tag{3}$$

If we know the tilts of the motion plane in the $xt$- and $yt$-manifolds, i.e.

$$\tan\gamma_1 = \frac{k_x}{k_t} \qquad \text{and} \qquad \tan\gamma_2 = \frac{k_y}{k_t} \tag{4}$$

we can determine the normal velocity, $\mathbf{v}$. However, the normal of the motion plane, $\mathbf{k}$, is all that is needed to determine the normal velocity.

The tilts $\tan\gamma_1$ and $\tan\gamma_2$ can be estimated optimally in the Total Least Square, TLS, error sense as the local directions of the 2D lines in the $xy$- and $yt$-manifolds by using complex convolution, [2] and [1].

$$\tilde{u}_1 = \int\int (\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial t})^2 dxdt \tag{5}$$

$$\tilde{u}_2 = \int\int (\frac{\partial f}{\partial y} + i\frac{\partial f}{\partial t})^2 dydt \tag{6}$$

It is worth noting that these quantities $\tilde{u}_1$ and $\tilde{u}_2$ are complex valued and that the "tilde" denotes that these are TLS estimations of the true directions. Here $f$ is the 3D image sequence, but the integrations are carried out in 2D planes of the sequence. 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. To be precise, the complex numbers $\tilde{u}_1$ and $\tilde{u}_2$ estimate $2\gamma_1$ and $2\gamma_2$ as follows

$$\tilde{u}_1 = m_1(\cos(2\gamma_1) + i\sin(2\gamma_1)) = m_1\exp(i2\gamma_1) \tag{7}$$

$$\tilde{u}_2 = m_2(\cos(2\gamma_2) + i\sin(2\gamma_2)) = m_2\exp(i2\gamma_2) \tag{8}$$

where $m_1$ and $m_2$ are certainty measures. In consequence, the arguments of $\tilde{u}_1$ and $\tilde{u}_1$, must be halved to yield the two tilt angles, $\gamma_1$ and $\gamma_2$ providing for an approximation of the velocity (3).

$$\tilde{v}_x = \frac{k_x}{k_t} = \tan\gamma_1 = \tan(\frac{1}{2}\arg(\tilde{u}_1)) \tag{9}$$

$$\tilde{v}_y = \frac{k_y}{k_t} = \tan\gamma_2 = \tan(\frac{1}{2}\arg(\tilde{u}_2)) \tag{10}$$

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

In our implementation we first used (5)-(6) to compute the two direction angle components needed to obtain the tilts, (9)-(10), which in turn enabled us to estimate the normal image velocities in lip images, via (3). In that, only processing along two planes embedded in 3D spatiotemporal images were needed.

We want to quantify the accuracy of this motion estimation. We do this by studying the results when the method is applied to synthetic image sequences where the velocities are known. Fig. 2a 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. 2b we show the profile along a line (indicated in (a)) where we can observe the spatial frequency in the test image. Fig. 2c illustrates the obtained optical flow estimation for one frame. The length of the arrows represent 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. This velocity direction accuracy is given further precision for the white circle in Fig. 2d,

[1] Note that the normal velocity in the $xy$-plane is 2D and it is invariant to a sign change of the gradient, i.e. the vectors $\nabla \mathbf{f}$ and -$\nabla \mathbf{f}$ represent the same plane, encoding the same velocity vector in the $xy$ plane


IEEE COMPUTER SOCIETY

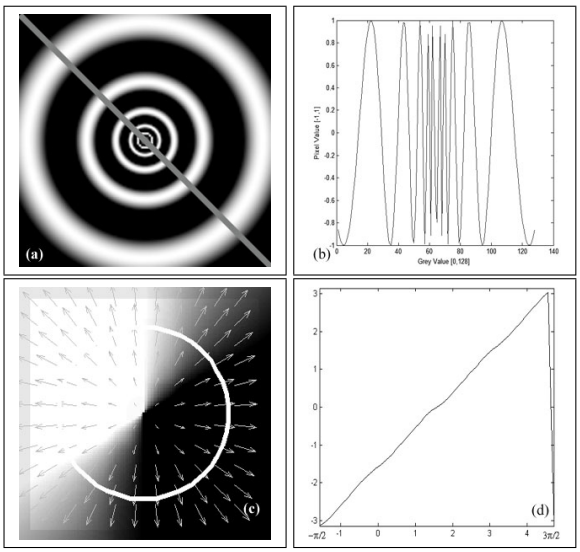


**Figure 2. a) Expanding waves test image. b) Profile of (a) along the indicated line. c) The estimated normal optical flow vectors with the orientation estimation in background. d) The graph shows the estimated argument of (c) along the indicated circle.**

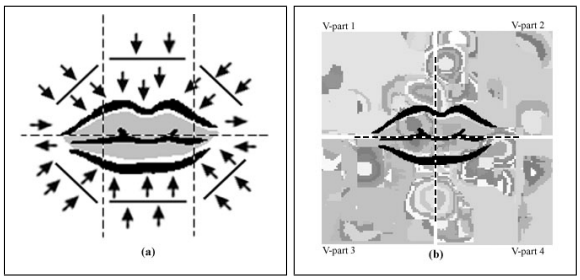


**Figure 3. The velocity vectors are divided into six regions, marked by dashed lines where each region is projected into a spatial direction marked by solid line. b) Shows the results of a clustering of the estimated velocity vectors that were divided into four parts by the dashed lines. The gray-values encode the absolute speeds in predefined directions.**

where we observe that the estimated velocity direction follows the true velocity direction very closely since the graph is linear. Additionally, the absolute speeds increase radially in agreement with the ground truth.

## 3. Experimental pre-processing and evaluation

Parameterization is carried out for finding a respresentation of the speech signal and the audio signal uniqe for each speaker. These parameterization are then merged by a fusion method to be send into a Gaussian Mixture Model (GMM) for evaluation and test of speaker verification. The Hidden Markov Toolkit (HTK) was used to process speech files and to perform GMM analysis [10].

### 3.1. Visual features

In each mouth-region frame we have numerous points, here 128X128 pixels, with dense 2D velocity vectors. Our goal is 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 3 directions $(0°, 45°, -45°)$ as marked with solid lines in 6 regions, (Fig. 3a). The motion vectors within each such region become real scalars that take the signs $+$ or $-$ depending on which direction they move relative their expected directions, the marked solid lines. Then the next step is to quantize these scalar velocities from being allowed arbitrary real scalars to a more limited set of values, here 20. These quantized velocities are obtained from the data by applying an automatic clustering technique, the fuzzy c-means [18], at four regions of the mouth-region (Fig. 3b). The obtained cluster-centers, and their corresponding cluster-populations, were used as a feature vector for each of the 4 regions. In consequence, each of these sub-regions had a 40 dimensional feature vector, consisting of 20 cluster-centers and 20 cluster-populations, summarizing the statistics of lip-motion.

### 3.2. Speech features

Speech signals are recorded at a 16 kHz samplings rate, and a speech frame with the length of 25 ms is extracted at every 10 ms. Frpm each speech frame a 39-dimensional acoustic parameters is carried out: 12-dimensional Mel-frequency cepstral coefficients with normalized log energy and 13 delta coefficients and 13 delta-delta coefficients.

### 3.3. Fusion

In a later step, the acoustic and visual features are merged into a single audio-visual feature vector. This allows us to develop a joint audio-visual model for person specific information in the data. 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 factor four is


IEEE COMPUTER SOCIETY

| Set / System | Evaluation | Test |
|---|---|---|
| Acoustic | 96% | 94% |
| Visual | 80% | 78% |
| Audio-Visual | 99% | 98% |

**Table 1. Verification results on evaluation and test set**

the rate factor between the audio and the visual data to the effect that we merge each of the four feature vectors of a visual frame with its own audio feature vectors. Accordingly, the merged feature vectors come at the rate of the audio feature vectors and have 79 elements, 39 audio and 40 lip-motion. The lip-motion information originating from the same instant are thus distributed in four consecutive samples of the merged data.

### 3.4. XM2VTS database and GMM

We have divided the XM2VTS audio-visual database [7] into 3 sets, according to Lausanne protocol [6] configuration I: 200 speakers were used as training set, 70 speakers as impostors for test and 25 speakers as impostors for evaluation. To investigate the speaker verification performance of the features within a GMM [8] framework using HTK toolbox, the following experiments were conducted on all speakers. The false acceptance rate FA and the false rejection rate FR were calculated.

### 3.5. Results

In the first experiment we used a GMM based speech recognition system based only on acoustic features. The threshold function, obtained from the evaluation set, is used in the test set to map the score into the confidence interval $[0, 1]$. From the verification results on the test set, we obtain the equal error rate (EER) of 6%. The verification rate of 295 speakers was thus 94%. In the second test, the system used merged visual and acoustic features on the same basis as the earlier system. The verification rate of the whole database was 98% , where the threshold from the evaluation set was utilized.

To quantify the biometric verification power of the visual features, we carried out the same experiments with these features alone. The verification performance using the threshold of the evaluation set was 78%. In Table. 1 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 give 98% correct classification which is 3-4% better than audio based speaker verification. This result shows that the audio-visual system achieves better performance than the audio-only system. It is worth noting that the result of the recognition rate in speaker verification is already high (94%) which means that any improvement is difficult. The verification results further show the importance of the visual signal as a complementary information source.

## 4. Conclusions

It is experimentally verified that the problem of velocity estimation in an image described by a local intensity function is possible to solve as a part of symmetry description of a neighborhood projected from 3D to a 2D planes. The presented results indicate that orientation of projected 2D image give the normal of an optimal plane which estimate velocities with sufficient accuracy to complement a speaker verification system. This solution is computationally efficient alternative to the optical flow computations based on 3D eigenvalue analysis.

Visual features are extracted from face image sequences to encode the motion statistics of the lips. Considering the size of test database, the performance of the system supports the conclusion that lip information is an important cue for voice or speech related person identification.

## References


[1] J. Bigun. *Vision with direction*. Springer, Heidelberg, Halmstad, SWEDEN, 2006.

[2] J. Bigun, G. Granlund, and J. Wiklund. Multidimensional orientation estimation with applications to texture analysis of optical flow. *IEEE-Trans Pattern Analysis and Machine Intelligence*, 13(8):775–790, 1991.

[3] K. R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

[4] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 1206*, pages 319–326, 1997.

[5] J. Luettin and G. Maitre. Evaluation protocol for the extended m2vts database $xm2vtsdb$. 1998. in: IDIAP Communication 98-054, Technical report R R-21, number = IDIAP.

[6] K. Messer, J. Matas, J. Kittler, and J. Luettin. Xm2vtsdb: The extended m2vts database. *In Second International Conference of Audio and Video-based Biometric Person Authentication*$ICSLP'96$, pages 72–77, 1999.

[7] D. Reynolds and R. Rose. Robust text-independent speaker identification using gaussian mixture models. *IEEE transactions on Speech and Audio processing* $ICASSP'90$, 3(1):72–83, 1995.

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