# Lip biometrics for digit recognition

Maycel Isaac Faraj and Josef Bigun

Halmstad University, 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 presents a speaker-independent audio-visual digit recognition system that utilizes speech and visual lip signals. The extracted visual features are based on line-motion estimation obtained from video sequences with low resolution (128 ×128 pixels) to increase the robustness of audio recognition. The core experiments investigate lip motion biometrics as stand-alone as well as merged modality in speech recognition system. It uses Support Vector Machines, showing favourable experimental results with digit recognition featuring 83% to 100% on the XM2VTS database depending on the amount of available visual information.

## 1 Introduction

Automated speaker and speech recognition have been suggested combining visual features to improve the recognition rate in acoustically noisy environments [1][2][3][4][5]. Dealing with digit recognition only, parts of this work are reported [6] which is related to [7][8]. The later is concerned with person authentication. A merger of this work and [6] as a journal article is under preparation. The reports [9][10][11][12][13] suggest visual features based on the shape and intensity of the lip region due to the changes in the mouth shape including the lips and tongue. The dynamic visual lip features carry significant phoneme-discrimination information embedded in motion information which can be modelled by moving-line patterns also known as *normal image velocity* [7][14], with no requirement of iterative process on detecting the mouth contour.

We exploit direct feature fusion to obtain the audio-visual observation vectors by concatenating the audio and visual features. The fused feature sequences are then modelled with a Support Vector Machine (SVM) classifier for digit recognition.

## 2 Feature extraction

### 2.1 Visual features

Bigun et al. proposed a motion estimation technique based on multidimensional structure tensor by solving eigenvalue problem [15], e.g. allowing the minimization process of fitting a line or a plane to be carried without the Fourier Transform. However, this method can be excessive for applications that only need

line-motion features. We assume that the local neighbourhood in the lip image contains parallel lines or edges which is a realistic assumption [8].

Lines translated with a certain velocity in the spatio-temporal image will generate planes with a normal, where the normal is estimated by using the total-least-square-error (TLS) [15]. We denote the normal unit of the plane as $\mathbf{k} = (k_x, k_y, k_t)^T$ and the projection of $\mathbf{k}$ represents the direction vector of the line's motion. The velocity vector can be written as follows.

$$\mathbf{v} = v\mathbf{a} = -\frac{k_t}{k_x^2 + k_y^2}(k_x, k_y)^T = -\frac{1}{(\frac{k_x}{k_t})^2 + (\frac{k_y}{k_t})^2}\left(\frac{k_x}{k_t}, \frac{k_y}{k_t}\right)^T, \tag{1}$$

where $v$ is the absolute speed in normal the direction and $\mathbf{a}$ is the direction of the velocity and $\mathbf{k} = (k_x, k_y, k_t)$ is the normal. The normal velocity estimation problem becomes a problem of solving the tilts ($\tan\gamma_1 = \frac{k_x}{k_t}$) and ($\tan\gamma_2 = \frac{k_y}{k_t}$) of the motion plane in the $xt$ and $yt$ manifolds. The estimated velocity components can by simple geometry be written as follows.

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

$$\frac{k_y}{k_t} = \tan\gamma_2 = \tan(\frac{1}{2}\arg(\tilde{u}_2)) \Rightarrow \tilde{v}_y = \frac{\tan\gamma_2}{\tan^2\gamma_1 + \tan^2\gamma_2} \tag{3}$$

where the arguments of $\tilde{u}_1$ and $\tilde{u}_2$ represent the TLS estimations of $\gamma_1$ and $\gamma_2$ in the local 2D manifolds $xt$ and $yt$ respectively, but in the double angle representation [16]. The tilde over $v_x$ and $v_y$ denote that these quantities are estimations of $v_x$ and $v_y$.

We have dense 2D-velocity vectors, $(v_x, v_y)^T$, in each mouth-region frame (128×128 pixels). The 2D velocity feature vectors $(v_x, v_y)^T$ at each pixel are reduced to 1D scalars with a certain sign + or − depending on which direction they move relative to their expected spatial directions 0°, 45°, −45° – marked with 3 different greyscale shades in 6 regions in **Fig. 1**.

$$f(p,q) = \|(v_x(p,q), v_y(p,q))\| * sgn(\angle(v_x(p,q), v_y(p,q))),\ p,q = 0\ldots127. \tag{4}$$

Next, we want to quantize the estimated velocities from arbitrary real scalars to a more limited set of values. The quantized speeds are obtained from the data by applying a mean approximation as follows.

$$g(l,k) = \sum_{p,q=0}^{N} f(Nl+q, Nk+p),\ \ p,q = 0\ldots(N-1),\ l,k = 0\ldots(M-1) \tag{5}$$

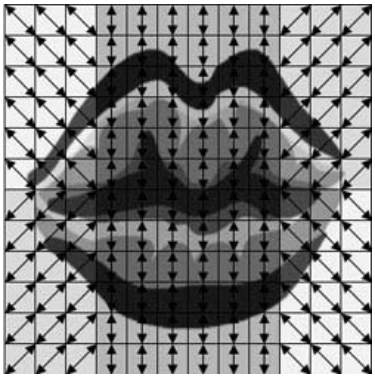

**Fig. 1.** Illustration of velocity estimation quantification and reduction.

where $N$ and $M$ represent the window size of the boxes (**Fig. 1**) and the number of boxes, respectively. The visual feature dimension of the lip-motion are represented by 144-dimensional ($M \times M$) feature vectors, whereas the original dimension before reduction is $128 \times 128 \times 2 = 32768$.

### 2.2 Acoustic features

We use the Mel-Frequency Cepstral Coefficient (MFCC) to extract acoustical feature information [17]. The MFCC speech features were generated by the Hidden Markov Model Toolkit (HTK) [18] processing the audio-data stream. The MFCC feature vector dimension is 39, containing 12 cepstral coefficients with normalized log energy, 13 delta coefficients (velocity), and 13 delta-delta coefficients (acceleration).

## 3 Classification by Support Vector Machine

Support Vector Machine (SVM) is a discrimination-based binary method using a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ as a decision boundary between two or several classes. For linearly separable training dataset labelled pairs $\mathbf{x}_i, y_i, i = 1, \ldots, l$, where $\mathbf{x}_i \in \Re^n$ and $\mathbf{y} \in \{1,\text{-}1\}^l$, the following equation is verified for each observation data (here feature vectors).

$$d_i(w^T \mathbf{x_i} + b) \geq 1 - \xi_i \; for \; i = 1, 2, ..., l \; \xi_i > 0, \tag{6}$$

where $d_i$ is the label for sample data $\mathbf{x}_i$ which can be +1 or -1; $\mathbf{w}_i$ and $b$ are the weights and bias that describe the hyperplane. In our experiment we use the inner-product kernel function as RBF kernel along with *one against one* class approach, further details found in [6]. For all the tests we use the SVM toolkit [19].

| Word | Audio features | Visual features | Audio-Visual features |
|---|---|---|---|
| 0 | 89% | 70% | 92% |
| 1 | 90% | 77% | 100% |
| 2 | 86% | 60% | 89% |
| 3 | 90% | 75% | 96% |
| 4 | 89% | 55% | 85% |
| 5 | 90% | 50% | 83% |
| 6 | 100% | 90% | 100% |
| 7 | 93% | 100% | 100% |
| 8 | 91% | 54% | 83% |
| 9 | 90% | 49% | 85% |

**Table 1.** Digit-recognition rate of all digits using protocol 2 in one against one SVM.

## 4 Experimental results

### 4.1 XM2VTS database

The experiments in this paper are conducted by the XM2VTS database [20], using the sentence "0 1 2 3 4 5 6 7 8 9" for all 4 sessions. It is of importance to note that the XM2VTS visual data is difficult to use as it is for digit recognition experiments because neither the audio-speech nor the visual-lip data are segmented. For each speaker of the XM2VTS database, the utterance " 0 1 2 3 4 5 6 7 8 9" was semi-automatically segmented into single-digit sub sequences 0 to 9, further details can be found in [6]. For all our tests we used *protocol 2* defined in [6] because XM2VTS is not designed for digit recognition but person authentication. However, it is the largest audio-visual database that is available, currently.

### 4.2 Digit recognition

Digit recognition results for all the systems based on only acoustic, only visual and merged audio visual feature information are presented in Table **1**. The best digit recognition rates are obtained for *1*, *6*, and *7* by 100%, whereas lower digit recognition rates are obtained for *4*, *5*, *8*, and *9*. One notable cause why the results in Table **1** vary is the lack of data (especially visual information) for certain digits. We could verify during the experiments that when uttering the words zero to nine in a sequence without silence between words the amount of visual data is notably less in the words *4*, *5*, *8*, and *9* in comparison to other digits in the XM2VTS database. Moreover, the amount of audio-visual speech for each digit is dependent on the manner and the speed of the speaker. The average of the digit recognition over all digits is $\approx$ 68% and $\approx$ 90% for only visual and only audio system respectively.

## 5 Conclusion and discussion

We present a digit recognition system exploiting lip motion information in dynamic image sequences using a large number of speakers with no use of iterative processing or assumption on successful lip-contour tracking. The presented experimental results confirm the importance of adding visual lip movement information in digit-recognition systems. Improvements of digit recognition utilizing our motion features to achieve higher recognition performance than audio-alone are verified.

An important matter is that the utterance of *4*, *5*, *8*, and *9* contain less audio and visual-lip information. The poor recognition performance of these digits along with manual inspection are indicators of that XM2VTS database do not contain sufficient amounts of visual information on lip movements.

## References


1. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.: Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE **91**(9) (2003) 1306–1326
2. Brunelli, K.R., Falavigna, D.: Person identification using multiple cues. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(10) (1995) 955–966
3. Chibelushi, C., Deravi, F., Mason, J.: A review of speech-based bimodal recognition. IEEE Transactions on Multimedia **4**(1) (2002) 23–37
4. Duc, B., Fischer, S., Bigun, J.: Face authentication with sparse grid gabor information. IEEE International Conference Acoustics, Speech, and Signal Processing **4**(21) (1997) 3053–3056
5. Tang, X., Li, X.: Video based face recognition using multiple classifiers. Sixth IEEE International Conference on Automatic Face and Gesture Recognition *FGR*2004 - IEEE Computer Society (2004) 345–349
6. Faraj, M.I., Bigun, J.: Speaker and speech recognition by audio-visual lip biometrics. The 2nd International Conference on Biometrics, Seoul Korea, 2007 (2007)
7. Faraj, M.I., Bigun, J.: Person verification by lip-motion. 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2006) 37–45
8. Faraj, M.I., Bigun, J.: Audio-visual person authentication using lip-motion from orientation maps. Article accepted for publication in Pattern Recognition Letters – 2007 (2007)
9. Luettin, J., Maitre, G.: Evaluation protocol for the extended m2vts database *xm2vtsdb*. (1998) in: IDIAP Communication 98-054, Technical report R R-21, number = IDIAP - 1998.
10. Dieckmann, U., Plankensteiner, P., Wagner, T.: Acoustic-labial speaker verification. Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 1206 (1997) 301–310
11. Jourlin, P., Luettin, J., Genoud, D., Wassner, H.: Acoustic-labial speaker verification. Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 1206 (1997) 319–326
12. Chen, T.: Audiovisual speech processing. IEEE Signal Processing Magazine **18**(1) (2001) 9–21

13. Liang, L., Zhao, X.L.Y., Pi, X., Nefian, A.: Speaker independent audio-visual continuous speech recognition. IEEE International Conference on Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 **2** (2002) 26–29
14. Kollreider, K., Fronthaler, H., Bigun, J.: Evaluating liveness by face images and the structure tensor. In AutoID 2005: Fourth Workshop on Automatic Identification Advanced Technologies - IEEE Computer Society (2005) 75–80
15. Bigun, J., Granlund, G., Wiklund, J.: Multidimensional orientation estimation with applications to texture analysis of optical flow. IEEE-Trans Pattern Analysis and Machine Intelligence **13**(8) (1991) 775–790
16. Granlund, G.H.: In search of a general picture processing operator. Computer Graphics and Image Processing **8**(2) (1978) 155–173
17. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on Acoustics, Speech, and Signal Processing **28**(4) (1980) 357–366
18. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The htk book (for htk version 3.0). (2000) http://htk.eng.cam.ac.uk/docs/docs.shtml.
19. Chang, C.C., Lin, C.J.: Libsvm–a library for support vector machines. software available at www.csie.ntu.edu.tw/ cjlin/libsvm (2001)
20. Messer, K., Matas, J., Kittler, J., Luettin, J.: Xm2vtsdb: The extended m2vts database. In Second International Conference of Audio and Video-based Biometric Person Authentication, $ICSLP'96$ (1999) 72–77