Viseme Classification for Talking Head Application

Mariusz Leszczynski and Władysław Skarbek

Faculty of Electronics and Information Technology
Warsaw University of Technology
W.Skarbek@ire.pw.edu.pl

Abstract. Real time classification algorithms are presented for visual mouth appearances (visemes) which correspond to phonemes and their speech contexts. They are used at the design of *talking head application.* Two feature extraction procedures were verified. The first one is based on the normalized triangle mesh covering mouth area and the color image texture vector indexed by barycentric coordinates. The second procedure performs Discrete Fourier Transform on the image rectangle including mouth w.r.t. a small block of DFT coefficients. The classifier has been designed by the optimized LDA method which uses two singular subspace approach. Despite of higher computational complexity (about three milliseconds per video frame on Pentium IV 3.2GHz), the DFT+LDA approach has practical advantages over MESH+LDA classifier. Firstly, it is better in recognition rate more than two percent (97.2% versus 99.3%). Secondly, the automatic identification of the covering mouth rectangle is more robust than the automatic identification of the covering mouth triangle mesh.

1 Introduction

This research refers to a development of software tools supporting animation of human face models integrated with Polish speech generator.

With a gradual performance progress of computer systems w.r.t. computing and transmission speed the *talking head applications* show higher realism in speech and dynamic visual face appearance (viseme).

Except the performance of speech generator, the synchronization between the spoken content and facial *visual content,* is of high importance. The visual content should not only provide the time correspondence of face image and related sound but also respect the semantic context of the speech, and the internal emotions of the speaker.

One of the main tasks in *talking head system* is the design of a correspondence table between visemes and phonemes (CTVP table). This correspondence is of *one to many* relational type. We can convert this relation to a mapping if we consider a *speech context* for the particular phoneme. In practice to get a unique viseme to speech context, it is enough to take into account three phonemes for such context: the current phoneme, the previous one, and the next one.

A. Gagalowicz and W. Philips (Eds.): CAIP 2005, LNCS 3691, pp. 773–780, 2005.

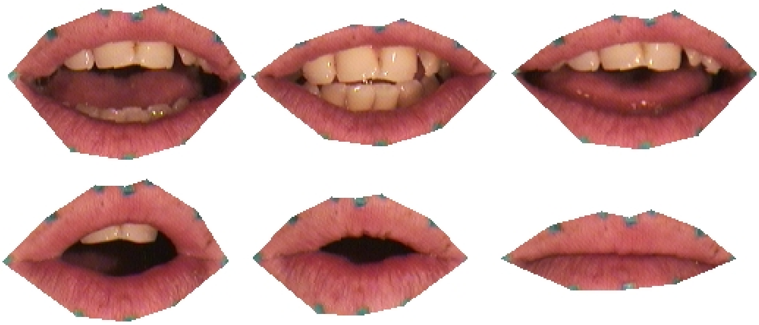


Fig. 1. Representative images for six major viseme classes – the 16 minor classes are obtained by discrimination between small, medium, and high degree of mouth opening within the first five major classes

In case of Polish speech patterns stored in the CORPORA database [2], the design of *phoneme context to viseme mapping* requires recording of video and audio material lasting about 1000 seconds. Therefore we get more than 25000 visemes to be classified and assigned to recognized phonemes context. This amount excludes manual implementation. Both, an automatic viseme classifier and phoneme classifier are necessary to complete the design of CTVP table.

For the phoneme classifier we have used a speech recognition engine based on HTK toolkit (cf. [7]). As a side effect the speech recognition program produces the phoneme and diphone transcription labelled by time information. Having such timing we could segment the video sequence into phoneme related groups. From each group this video frame was selected for viseme classification which was closest in time to the middle of phoneme time interval, i.e. to the beginning of diphone interval. The recognized viseme class (cf. Fig.1) was joined to the phoneme context list. At the end, from each phoneme list the class id was selected using the majority rule.

This work explains how the viseme classifier had been designed to support the creation of CTVP table. To this goal the classification performance of 80% could be sufficient. However, we are going to use our viseme classifier to animate the human head model on the basis of live video. Therefore the real time and the high performance of the classifier are the main objectives of our research.

2 Image Normalization

The realistic visual speech can be achieved by integrating the person specific face model with mouth model optionally augmented with the model of chin and cheeks. Using a triangle mesh (cf. Fig.2), we can cover those speech sensitive areas and try to get the model for at least two goals: viseme classification and mouth animation.

Alternatively we can approximate the mouth area by a least rectangle touching lips from outside (cf. Fig.3 upper part). Obviously, the triangle mesh ap-

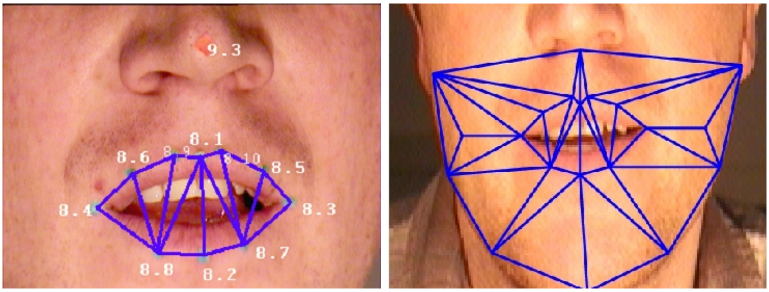


Fig. 2. Triangle mesh for mouth with MPEG-4 FAP points depicted (left), and its neighbourhood (right)

proximation of mouth area is more exact than rectangular one and therefore a texture vector built from the rectangle includes components hard for matching. In this case change to 2D Discrete Fourier Transform (DFT) domain enables correct matching of mouth images normalized to reference mouth rectangle. As the vertical variability of the mouth image dominates the horizontal one, we expect that out of three corner blocks (cf. Fig.3 lower part) in DFT domain (usually considered at DFT based feature extraction) only the one corresponding to the least frequencies (without conjugated part) will be important for classification. Our expectation has been confirmed by the experiments.

In mesh approach we deal with variations of the mesh shape and of the mesh texture (*appearance*). In order to make comparable two meshes we have to normalize them with respect to a reference mesh.

We perform the nonlinear normalization of the mesh by mapping each triangle in the current image onto the corresponding triangle in the reference im-

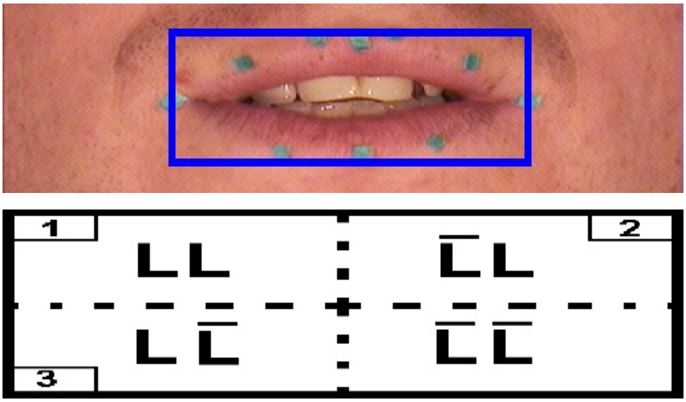


Fig. 3. The rectangle including mouth area (upper), and channel subdivision for 2D DFT (lower)

age.Each local mapping is affine, but globally we obtain the mapping which is piecewise affine.

Let the i-th triangle $\Delta_i(P_0, P_1, P_2)$ in the reference mesh $\mathcal{M}$ be mapped by the affine mapping $A_i(P) = B_i P + t_i$ onto the triangle $\Delta'_i(P'_0, P'_1, P'_2)$ in the current mesh $\mathcal{M}'$, where B_i is the square matrix, t_i is the vector, $P \in \Delta_i$, $P' \in \Delta'_i$, $i = 1, \ldots, K$. Then we have the following properties:

1. The piecewise affine mappings $A_1, \ldots, A_K$ are *continuous mappings* of $\mathcal{M}$ onto $\mathcal{M}'$ in geometric space
2. If $P = \alpha_0 P_0 + \alpha_1 P_1 + \alpha_2 P_2$ has the barycentric coordinates $\alpha_0, \alpha_1, \alpha_2$ w.r.t. the triangle $\Delta_i(P_0, P_1, P_2)$ then the point $A_i(P) = \alpha_0 P'_0 + \alpha_1 P'_1 + \alpha_2 P'_2$, i.e. it has *the same barycentric coordinates* with respect to the triangle $\Delta'_i(P'_0, P'_1, P'_2)$:

$$\begin{aligned} A_i(P) &= B_i P + t_i = B_i(\alpha_0 P_0 + \alpha_1 P_1 + \alpha_2 P_2) + (\alpha_0 P_0 + \alpha_1 P_1 + \alpha_2)t \\ &= \alpha_0(B_i P_0 + t) + \alpha_1(B_i P_1 + t) + \alpha_2(B_i P_2 + t) = \alpha_0 P'_0 + \alpha_1 P'_1 + \alpha_2 P'_2 \end{aligned}$$

3. If $f' : \Delta'_i(P'_0, P'_1, P'_2) \rightarrow \mathcal{C}_{RGB}$ is the texture mapping in the current mesh then the mapping $f : \Delta_i(P_0, P_1, P_2) \rightarrow \mathcal{C}_{RGB}$ is defined by the barycentric coordinates for $i = 1, \ldots, K$ as follows:

$$f(P) = f(\alpha_0 P_0 + \alpha_1 P_1 + \alpha_2 P_2) \triangleq f'(\alpha_0 P'_0 + \alpha_1 P'_1 + \alpha_2 P'_2) \tag{1}$$

The above substitution transfers the texture from the current mesh onto the reference mesh with possible deformation of linear segments which intersect at least two triangles in the mesh.

3 LDA for Mouth Classification

The advantage of having all texture classes (in mesh case) or DFT coefficients classes (in rectangular case) in common space $\mathbb{R}^N$ allows us to use the Linear Discriminant Analysis (LDA) to design the extremely fast classifier of linear complexity $O(N)$.

Before we reached LDA feature vector of dimension five, the general Fisher LDA criterium (cf. [3,5,6]) had been used for K dimensional training feature vector $y_i = W^t x_i$, $x_i \in \mathbb{R}^N$, $i = 1, \ldots, L$, $y \in \mathbb{R}^K$, $W \in \mathbb{R}^{N \times K}$:

$$W_{opt} = \arg\max \frac{\text{between class variance for } \{y_i\}}{\text{within class variance for } \{y_i\}} = \frac{tr(W^t S_b W)}{tr(W^t S_w W)} \tag{2}$$

where S_b, S_w are the between and within class scatter matrices.

The above criterium has points of singularity if W is arbitrary. Therefore Fisher imposed the following constraints on the domain of W :

$$W^t S_w W = I,\ W \perp \ker(S_w) \tag{3}$$

This leads us to the following steps to obtain the optimal W described in details as two singular subspace method in [1] with tuning parameters q equal to the dimension of the intra-class singular subspace (cf. [4]):

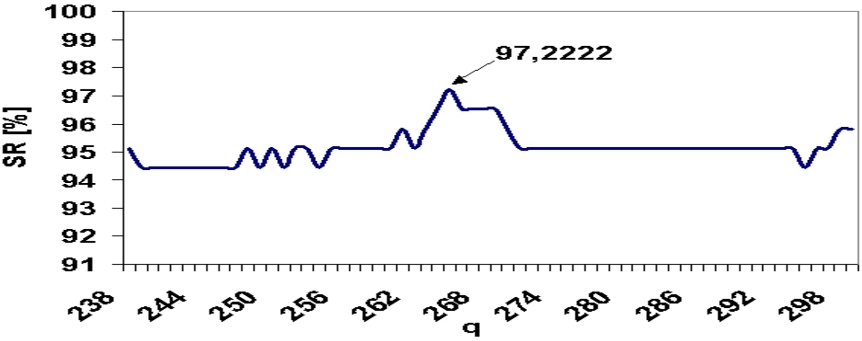
1. Class mean shifting of the training sequence: $X = [x_1, \ldots, x_L]$;
2. Grand mean shifting for class means: $M = [m_1, \ldots, m_C]$;
3. Singular Value Approximation for X with subspace dimension equal to q :

$$[U_q, \Sigma_q] := sva(X, q);\ A_q = U_q \Sigma_q^{-1};$$

4. Whitening of columns in $M : M = A_q^t M$;
5. Singular Value Approximation for M with subspace dimension equal to r :

$$V_r := sva(M, q);\ W = A_q V_r;$$

6. Return W;



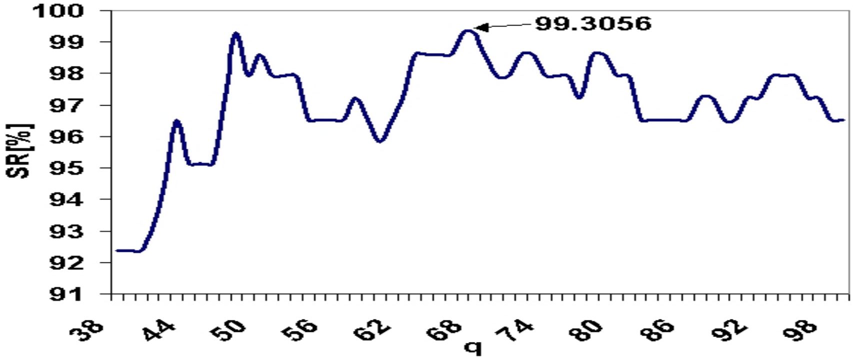


Fig. 4. Recognition rate versus LDA tuning parameter q when $r = 5$: for MESH+LDA (upper graph) and DFT+LDA (lower graph)

In case of mesh based feature vector (MESH+LDA) and DFT based feature vector (DFT+LDA), the Fig.4 shows the expected behavior of recognition rate versus the tuning parameter q.

The vector LDA features with maximum possible value $r = C - 1 = 5$ gives the best results.

The LDA feature $y = W^t x$ for the texture vector x is classified by the distance to LDA features $y_i = W^t x_i$ representing the mouth appearance classes $i = 1, \ldots, 6$:

$$i_{opt} = \arg \min_{1 \le i \le 6} \|y - y_i\|^2 \qquad (4)$$

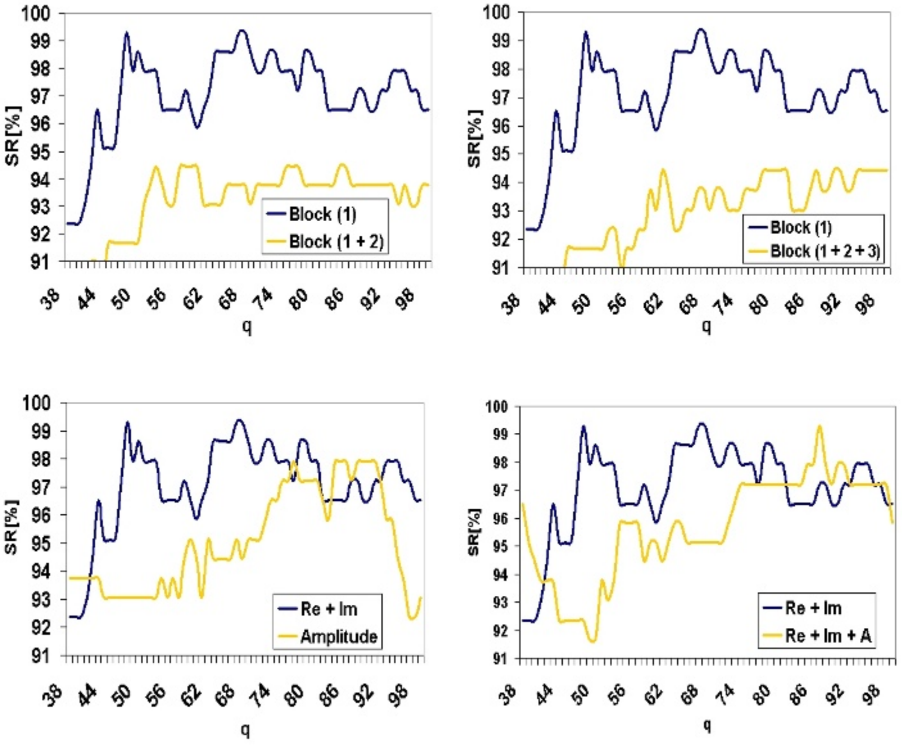


Fig. 5. Recognition rate versus LDA tuning parameter q when $r = 5$: for different choice of DFT channels (in upper graphs block 1 contains coefficients of LL frequencies, 2 – $\bar{L}L$, 3 – $L\bar{L}$) and different combinations of real, imaginary and amplitude parts in DFT (lower graphs)

4 Experimental Results

For the training of models for feature extraction, 497 mouth image were selected with unbalanced distribution in the classes what corresponds to the distribution in the whole recorded video sequence:

$$L_1 = 127,\ L_2 = 123,\ L_3 = 42, L_4 = 89, L_5 = 37,\ L_6 = 79$$

For the testing stage, 152 frames were selected independently of training frames.

In mesh texture case the best result (97.2% – cf. Fig.4 upper part) is achieved for lower resolution image with subsequent subsampling of texture vector. Since in case of LDA, the extraction time is independent of q, we accept higher values of q giving higher generalization of the classifier even if the recognition rate is slightly higher for lower values of q.

In rectangular DFT case the best recognition result (equal to 99.3%) is achieved for the following setup of parameters:

1. DFT block LL for horizontal frequencies $0-4$ and vertical frequencies $0-19$ (cf. Fig.3 at lower part and graphs of Fig.5 at upper part)
2. DC component is skipped
3. imaginary and real parts of all coefficients in blok LL are stacked in one vector of size 198 (contrary to the face classifier used in our system, the amplitude of DFT coefficients for mouth classifier has appeared to be insignificant – cf. graphs of Fig.5 at lower part)
4. intra-class singular subspace dimension equals to 67 (cf. Fig.4 lower part)
5. inter-class singular subspace dimension equals to 5

It appears that mouth images which were wrongly classified are only from the class of slightly opened mouth with visible upper teeth, without visible tongue. They were confused with opened mouth, visible upper teeth and visible tongue. However, by eye view (the important measure in talking head application) the difference between such two images is not annoying while watching the mouth animation.

5 Conclusion

Two real time algorithms MESH+LDA and DFT+LDA for visemes classification were compared.

Both algorithms benefit of optimization stage when the optimal first singular subspace dimension is selected in our LDA design. LDA matrix in mesh has about 30 times more elements than LDA matrix in DFT case. However this advantage at matrix computation is absorbed by dominating DFT computational time.

Preliminary feature extraction for MESH+LDA is slightly faster but less robust in case of automatic mesh identification.

DFT+LDA method is better than MESH+LDA in recognition rate more than two percent (97.2% versus 99.3%). Therefore for *talking head* applications, DFT+LDA technique is recommended.

Acknowledgments

The work presented was developed within VISNET, a European Network of Excellence (`http://www.visnet-noe.org`), funded under the European Commission IST FP6 programme.

References

1. Bober M., Kucharski K., and Skarbek W.: Face Recognition by Fisher and Scatter Linear Discriminant Analysis, in Computer Analysis of Images and Patterns, eds. Petkov N., Westenberg M., Springer LNCS 2756, 638:645, 2003
2. Grocholewski S.: CORPORA - Speech Database for Polish Diphones, 5th European Conference on Speech Communication and Technology EUROSPEECH '97 Rhodes, Greece, September 22-25, 1997
3. Fukunaga K.: Introduction to statistical pattern recognition (2nd ed). Academic Press, Boston, 1990
4. Golub G., Van Loan C.: Matrix Computations. Baltimore: Johns Hopkins University Press, 1996
5. Ripley B.D.: Pattern Recognition and Neural Networks. Cambridge University Press, 1996
6. Swets D.L., Weng J.: Using Discriminant Eigenfeatures for Image Retrieval, IEEE Trans. on PAMI, 18(8):831-837, August 1996
7. The Hidden Markov Model Toolkit (HTK) http://htk.eng.cam.ac.uk