Dictionary-Based Lip Reading Classification

Dahai Yu*, Ovidiu Ghita°, Alistair Sutherland*, Paul F. Whelan°
*School of Computing, °Vision Systems Group, School of Electronic Engineering
Dublin City University, Dublin 9, Ireland
dahai.yu2@mail.dcu.ie

Abstract

Visual lip reading recognition is an essential stage in many multimedia systems such as “Audio Visual Speech Recognition” [6], “Mobile Phone Visual System for deaf people”, “Sign Language Recognition System”, etc. The use of lip visual features to help audio or hand recognition is appropriate because this information is robust to acoustic noise. In this paper, we describe our work towards developing a robust technique for lip reading classification that extracts the lips in a colour image by using EMPCA feature extraction and k-nearest-neighbor classification. In order to reduce the dimensionality of the feature space the lip motion is characterized by three templates that are modelled based on different mouth shapes: closed template, semi-closed template, and wide-open template. Our goal is to classify each image sequence based on the distribution of the three templates and group the words into different clusters. The words that form the database were grouped into three different clusters as follows: **group1**(‘I’, ‘high’, ‘lie’, ‘hard’, ‘card’, ‘bye’), **group2**(‘you, ‘owe’, ‘word’), **group3**(‘bird’).

Keywords: Lip Reading, Template Model, EMPCA, K-Nearest Neighbour Classification.

1 Introduction

An automatic system able to recognize a significant number of signs from a sign language requires hand recognition and lip recognition. Although the hand features represent the primary source of information, the lip visual features may be used to enhance the performance of the sign recognition system since they are robust against noise perturbation or similar-looking signs (e.g. old and on, go and come, etc). In this paper, we attempt to evaluate whether lip motion can be used as an additional cue to improve the performance of systems designed to recognize sign language. The identification of words based on lip movement is a difficult task because: 1. Lips are highly deformable, they vary in shape, colour, specularity, and in their relation to surrounding features across individuals; [1] 2. There is too much similarity in lip motion in similarly pronounced words. To address these problems we present a robust method to cluster the words based on a three-template model that is employed to capture the distribution of these templates in the image sequence. To evaluate this approach we generate a dictionary-based database that consists of a number of images associated with different words. The basic approach is depicted in the following figure:

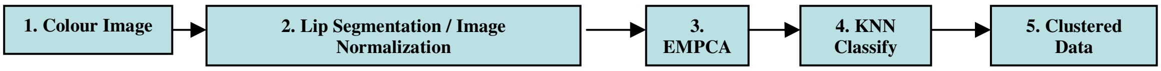


Figure 1: Outline of the developed approach for lip reading.

We have tested 7500 images generated by three subjects saying 15 different words. The aim of these tests is to evaluate the discrimination between words: ‘I’ vs. ‘you’, ‘I’ vs. ‘word’, ‘you’ vs. ‘bird’, etc. This paper is organized as follows. Section 2 introduces the lip detection and normalization procedure. The three templates model is presented in Section 3 and the classification scheme is detailed in Section 4. Experimental results are discussed in Section 5 while Section 6 concludes this paper.

2 Lip Detection & Image Normalization

There are many lip detection and tracking methods that have been proposed so far which include the use of a skin colour model, pseudo-hue and Active Contour Model [1,3,4]. The first component of the system performs lip segmentation by analyzing the pseudo-hue component of the colour image.

$$H(x,y)=R(x,y)/G(x,y)+R(x,y) \qquad (1)$$

Where *R* and *G* are the red and green components of the colour image and *H* is the pseudo-hue value.

The mouth region is detected based on the two corners of the upper lips and the mean flow is used to normalize image intensities as follows:

$$P(Rn,C1\ldots C30)=P(Rn,C1\ldots C30)/mean(P(Rn,C1\ldots C30)) \qquad (2)$$

Where P is the [40×30] normalised image, R is row and C is column, n is the row index (n=1… 40).

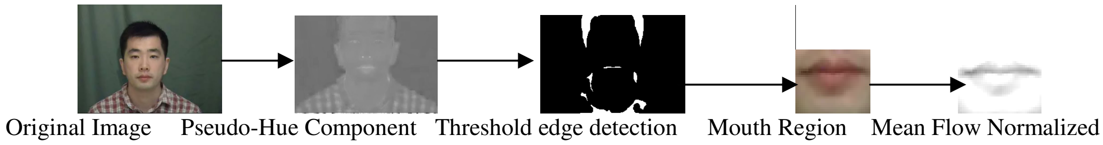


Figure 2: Mouth detection progress.

3 Three-Template Model

3.1 Template Model Definition

We introduce three templates based on mouth shapes that relate to particular phonemes. The relationship between phonemes and mouth shape is a many-to-one mapping [7]. For example, although 'o' and 'u' are acoustically distinguishable sounds, they are grouped into one template category as they have similar sequences of mouth shapes. At the current stage, all data used to construct the three templates is manually selected based on a visual examination of the mouth shapes that can represent particular phonemes. For instance, Fig. 3 shows the mouth shapes corresponding to different phonemes. The three templates are: **T1**-Closed template (standard mouth shape and mouth shapes that correspond to 'b' phoneme), **T2**-semi-closed template (mouth shapes that correspond to 'o' & 'u') and **T3**-wide open template (mouth shapes that correspond to 'I' & 'a'). We assume that the mouth always follows the sequence close/tight close, semi-open, open when speaking.

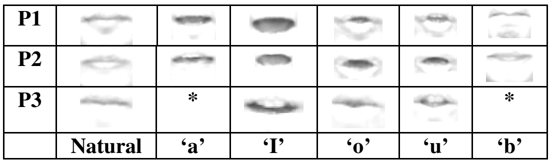

P1
P2
P3
*
*
Natural
'a'
'I'
'o'
'u'
'b'

Figure 3: The three templates employed to model particular phonemes. Note that there are no P3 templates that correspond to 'a' and 'b' mouth shapes.

3.2 Expectation-Maximization Principal Component Analysis

We employed the Expectation-Maximization (EM) PCA approach to identify the distribution of the three templates in the image sequences that define the words to be analyzed. EM-PCA has all the advantages of the EM algorithm in terms of estimating the maximum likelihood values for missing information directly at each iteration and is more appropriate to handle high dimensional data than standard PCA. It allows a simple and efficient computation of a few eigenvectors and eigenvalues to be extracted from large collections of high dimensional data [4]. The EM-PCA has two distinct stages the E-step and M-step:

$$\text{E-step: } W = (V^TV)^{-1}\, V^{-1}A; \qquad \text{M-step: } Vnew = AW^T(WW^T)^{-1}; \qquad (3)$$

Where ***'W'*** is the matrix of unknown states, ***'V'*** is test data vector, ***'A'*** is the observation data, $^{\mathbf{T}}$ is transpose operator. The image frame A is read as a matrix with m rows and n columns (m=40, n=30) and in equation 3 is converted into a one dimensional vector by reading the matrix in a raster scan mode as illustrated in Fig. 4.

$$\begin{pmatrix} A11, A12 \ldots A1n \\ A21, A22 \ldots A2n \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ Am1, Am2 \ \ldots Amn \end{pmatrix} \xrightarrow[\text{a one dimensional vector}]{\text{Conversion to}} \begin{bmatrix} A11, A12\ldots A1n, A21, A22\ldots. A2n, \ldots.., Am1, Am2, \ \ldots Amn \end{bmatrix}$$

Figure 4: Conversion of the matrix A to a one-dimensional vector.

For 7500 frames, there will be generated a [7500×1200] observation matrix and after the application of EMPCA the dimensionality of this data will be reduced to 100 dimensions [7500×100]. All this information is used for training and testing.

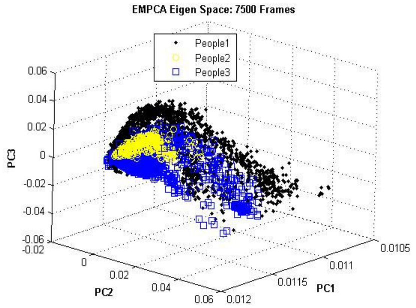


Figure 5: First three largest components plotted in the PCA space.
(Person 1: 4000 frames, Person 2: 3000 frames and Person 3: 500 frames)

4 Classification

4.1 Generating Template Model

All data used to construct the five different phonemes and standard mouth shapes are manually selected (see Section 3.1) and they are displayed in the PCA space. The diagrams of the largest three components associated with the three templates are illustrated in Fig. 6.

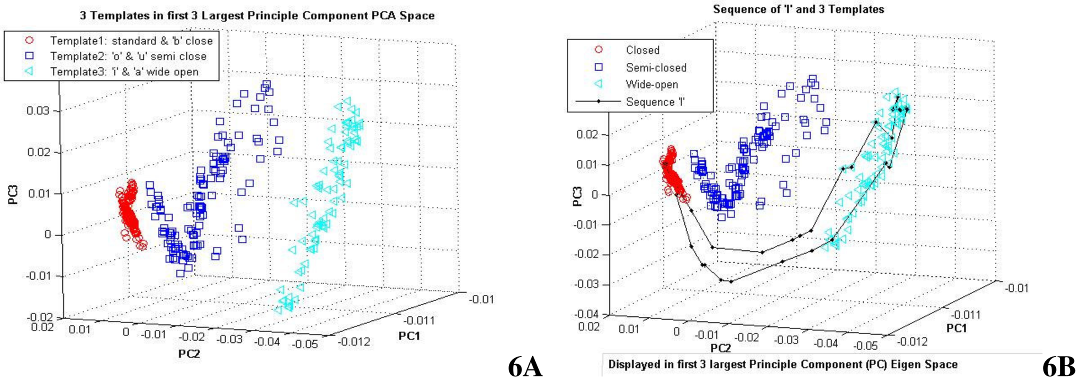


Figure 6: The largest three principal components associated with the three templates.
In 6B, the black line illustrates the sequence of word 'I', each black point is an image frame of the sequence.

4.2 KNN Classification

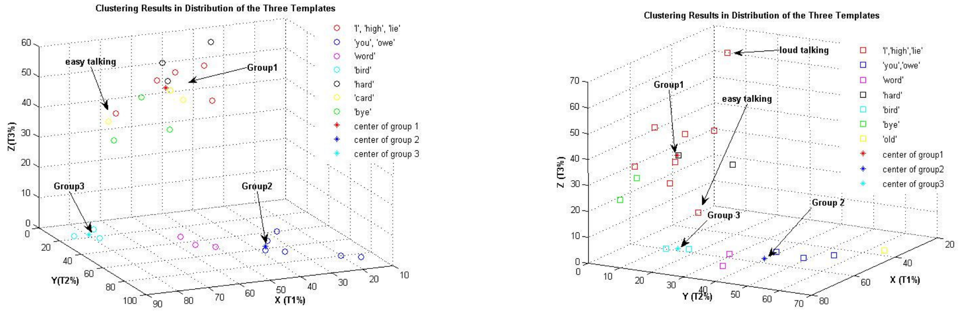
The nearest neighbour algorithm is a simple classification scheme where the input data is classified according to the distance to the nearest neighbour from some previously known classes [8]. If we have a sequence of images for one word, we use KNN to classify each image frame with respect to the three template models. Some frames are not classified to any template, so those frames are assigned as Not Classified Frames (NCF) (see Fig. 6B)

4.3 Clustering

For each word, the percentage of the total number of frames assigned to each of the three templates is calculated. Each word can be represented as a point in a 3D space whose axes represent the percentage result of the three templates in the image sequence. When all the words in the training set are projected onto this space, three distinct groups are visible as depicted in Fig. 7. During the test phase each new word is assigned to the nearest group based on the Euclidean distance between the data point and the centre of each group.

5 Experimental Results

Based on the distribution of the three templates in the image sequence, we can group the words in different clusters as illustrated in Fig. 7. The training results (average distribution value of the three templates for each group of words) and the classification rate are displayed in Table 1.



7A **7B**

Figure 7: Clustering results.

7A: 29 Sequences from Person 1(male); 7B: 21 Sequences from Person 2 (female)
X-Axis: percentage of closed template (**T1**); **Y**-Axis: percentage of semi-closed template (**T2**); **Z**-Axis: percentage of wide-open template (**T3**); Each point in Figure 7 defines the plot of the words contained in the database based on the occurrence of the three templates in the image sequence**. (*: T1+T2+T3+NCF=100%)**

	Person 1 (Male)				**Person 2 (Female)**			
Group	**T1**	**T2**	**T3**	**Classify Error**	**T1**	**T2**	**T3**	**Classify Error**
Group 1	41-65%	5-10%	35-47%	17%	40-54%	3-16%	22-43%	25%
Group 2	15-38%	51-85%	0-2%	33%	45-66%	33-54%	0-3%	17%
Group 3	75-84%	11-22%	0-5%	0%	62-73%	20-22%	2-6%	0%

Table 1: Classification results based on the distribution of the three templates
Group 1:'I', 'High', 'Lie', 'Bye', 'Hard', 'Card'; **Group 2:**'You', 'Owe', 'Word', 'Old'; **Group 3**: 'Bird'.

6 Conclusions

In this paper, we have proposed a robust lip reading method based on the overall distribution of three templates in the image sequence. The experimental results were encouraging and indicate that our method is feasible to group the words into distinct classes but is not able to robustly discriminate between words from the same group. Also during experimentation, it has been found that there is a large variation in mouth shapes for similar words generated by different persons. Additional errors were generated by the similar pronunciation of different words and by imperfect lip region extraction. We intend to further develop the approach presented in this paper in order to improve the classification results by increasing the number of templates that are used to characterize the fundamental mouth shapes (phonemes).

References

[1] N. Eveno, A. Caplier, P. Coulon (2004), "Accurate and Quasi-Automatic Lip Tracking, *IEEE Trans. Circuits Syst. Video Techn.* 14(5): 706-715.

[2] Y. L. Tian, T. Kanade (2000), "Robust lip tracking by combining shape colour and motion", *ACCV2000* (1)1040 -1045.

[3] A.V. Nefian, L.H. Liang, X. Liu, X. Pi, (2002) "Audio-Visual Speech Recognition", *Intel Technology & Research*. http://www.intel.com/technology/computing/applications/avcsr.htm

[4] S. Roweis (1998), "EM Algorithms for PCA and SPCA", *Advances in Neural Information Processing Systems* (10): 626-632.

[5] J.M. Keller, R. Gray, J.R. Givens (1985), "A Fuzzy K-nearest Neighbor Algorithm". *IEEE Trans. Systems Man Cybernet.* 15(4), 580-585.

[6] Z. Ghahramani, Machine Learning Toolbox, Version 1.0 01-04-96, copyright © University of Toronto.

[7] S.W. Foo, Y. Lian. (2004), "Recognition of visual speech elements using adaptively boosted HMM", *IEEE Transactions on Circuits and Systems For Video Technology*, 14(5) 693-705.

[8] J. Sim, S. Y. Kim, J. Lee (2005) **"**Prediction of protein solvent accessibility using fuzzy *k*-nearest neighbor method ", *Bioinformatics* 21(12):2844-2849.