A New Manifold Representation for Visual Speech Recognition

Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan

School of Computing & Electronic Engineering, Vision Systems Group
Dublin City University, Dublin 9, Ireland
Dahai.yu2@mail.dcu.ie

Abstract. In this paper, we propose a new manifold representation capable of being applied for visual speech recognition. In this regard, the real time input video data is compressed using Principal Component Analysis (PCA) and the low-dimensional points calculated for each frame define the manifolds. Since the number of frames that from the video sequence is dependent on the word complexity, in order to use these manifolds for visual speech classification it is required to re-sample them into a fixed number of keypoints that are used as input for classification. In this paper two classification schemes, namely the k Nearest Neighbour (kNN) algorithm that is used in conjunction with the two-stage PCA and Hidden-Markov-Model (HMM) classifier are evaluated. The classification results for a group of English words indicate that the proposed approach is able to produce accurate classification results.

Keywords: Visual speech recognition, PCA manifolds, spline interpolation, k-Nearest Neighbour, Hidden Markov Model.

1. Introduction

In recent years, visual speech recognition has became an active research topic and plays an essential role in the development of many multimedia systems such as audio-visual speech recognition (AVSR) [4], mobile phone applications and sign language recognition [10]. The inclusion of lip visual features as additional information in the development of audio or hand recognition algorithms has received interest from computer vision community because this information is robust to acoustic noise.

The aim of this paper is to detail the development of a visual speech recognition system that is able to achieve speech recognition using only the visual features that are extracted from the input video sequence. A review of the computer vision literature indicates that several approaches have been proposed to address the visual speech recognition based on the features that are extracted from the lips contour. In 1995, Luettin et al [6] applied Active Shape Models to identify the lips and extract the features that are used for visual speech recognition. In the same year, Bregler and Omohundro [18] proposed a different technique where the lip motions are encoded using nonlinear manifolds that are used to identify standard English phonemes. Later, Richard Harvey [17] proposed a new approach for speech recognition where the central part is a morphological transform called the sieve that is applied to calculate simple one-dimensional (1D) and two-dimensional (2D) measurements that are able to sample the lips shapes. The statistics calculated from these 1D and 2D measurements are concatenated into a feature vector that is used to train a standard HMM classifier. In 2002, Gordan et al [12] propose to apply a network of support

vector machines (SVM) classifiers for visual speech recognition [12]. This work was further advanced by Foo and Lian [9] and Dong et al [16] where adaptive boosting and HMM classifiers were applied to recognize visual speech elements. More recently, Yau et al [14] propose the use of image moments and multi-resolution wavelet images for visual speech recognition. In their approach, the input video data is represented by the motion history image that is decomposed by applying the discrete stationary wavelet transform.

From this short literature review we can conclude that most of the work was focused on the robust identification of small independent speech elements (called visemes [9]) while the word recognition is viewed as a simple combination between standard visemes. In visual speech process, a viseme (which is a mouth shape or a short sequence of mouth dynamics that are required to uniquely generate a phoneme or a group of phonemes in the visual domain) is regarded as the smallest unit that can be identified using visual information from the input video data. Although the words can be theoretically formed from a combination of standard visemes, in practice due to various pronunciation styles similar visemes can be associated with different visual signatures. In addition to this, the viseme identification within words is problematic since the transitions between consecutive visemes are not always easy to identify. In order to alleviate these problems, in this paper we formulate the visual speech recognition as the process of recognizing individual words based on a manifold representation. While this approach to visual speech recognition is appealing since provides a generic framework to recognise words without resorting to viseme identification, a few important issues need to be addressed. The first problem is to evaluate the discriminative power offered by the manifold representation while the second problem consists of designing the classification scheme that returns optimal results. These issues will be addressed in detail in the following sections of this paper.

This paper is organized as follows. Section 2 presents an overview of the proposed system while Section 3 describes the methodology applied to obtain the manifolds from input data. In Section 4 two classification schemes are detailed and their performance is analysed in detail in Section 5. Concluding remarks and discussions on the future work are provided in Section 6.

2 System Overview

The developed system for visual speech recognition consists of three main steps. In the first step, the lips are extracted from input video data. In order to achieve this goal we calculate the pesudo-hue [3] from the RGB components and the lips are segmented by applying a histogram based thresholding scheme (see Fig. 1). The aim of the second component is to generate the Expectation Maximization PCA (EM-PCA) manifolds and perform manifold interpolation and re-sampling. The role of the third component is to classify the manifolds calculated for the input speech sequence into a number of words that are contained in a database. For this implementation two classification schemes are evaluated, namely the second stage EM-PCA used in conjunction with a kNN classifier and the HMM classifier. The block diagram of the proposed visual speech recognition system is depicted in Fig. 2.

(a) (b) (c) (d) (e)

Fig. 1. Lip detection algorithm. (a) RGB image; (b) Pseudo-hue image; (c) Image resulting after the application of the histogram-based thresholding; (d) Lips region –pseudo hue; (e) Lips region – grayscale.

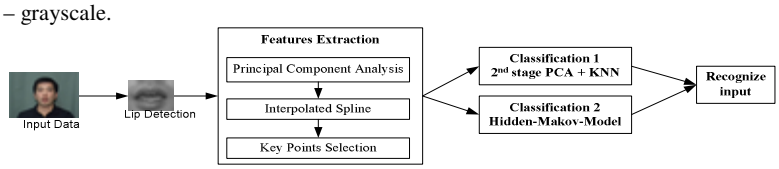


Fig. 2. Block diagram of the proposed visual speech recognition system: Step (1) Lip Detection from the pseudo-hue component; Step (2) Features Extraction; Step (3) Classification.

3 Manifold Representation

3.1 EM-PCA Mathematical background

Expectation-Maximization PCA (EM-PCA) is an extension of the standard PCA technique by incorporating the advantages of the EM algorithm in terms of estimating the maximum likelihood values for missing information. This technique has been originally developed by Roweis [7] and its main advantage over the standard PCA is the fact that it is more appropriate to handle large high dimensional datasets especially when dealing with sparse training sets. The EM-PCA procedure has two distinct stages, the E-step and M-step:

$$\text{E-step: } W = (V^T V)^{-1} V^{-1} A; \quad \text{M-step: } V_{new} = A W^T (W W^T)^{-1} \tag{1}$$

where ***'W'*** is the matrix of unknown states, ***'V'*** is the test data vector, ***'A'*** is the observation data matrix, and ^{T} defines the transpose operator.

3.2 Manifold calculation from input data

Human lips are highly deformable objects and it can be noticed that during the speech process they show variations in shape, color, reflection and also their relation to surrounding features such as tongue and teeth is very complex [5]. For visual speech recognition purposes, we would like to extract the information associated with the lip motions from the frames that define the input video sequence. As indicated in previous section, the lips are segmented in each frame by thresholding the pseudo-hue component calculated from RBG data and we encode the appearance of the lips in each frame as a point in a feature space that is obtained by projecting the input data onto the low dimensional space generated by the EM-PCA procedure. The feature points obtained after the projection on the low-dimensional space are joined by a plotline based on the frame order and in this way we generate a surface in the feature space that is called manifold [18]. In other words, the image region surrounding the lips in each frame (see Fig. 1e) is compressed using the EM-PCA technique that is detailed in Section 3.1. Since the manifolds encode the lips motion through image compression, the shape of the manifolds will be strongly related with the words spoken by the speaker and recorded in the input video sequence. In Fig. 3 is illustrated

the variation between three independent image sequences of the same word in the EM-PCA feature space. It can be noted that the shapes of the manifolds are very similar and can be interpreted as a word "signature".

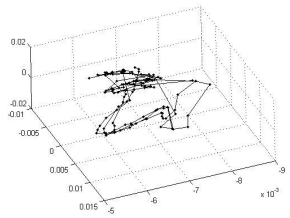


Fig. 3. Manifolds generated from three image sequences representing a same word. The appearances of the manifolds indicate that their shapes are similar and contain information in regard to the word spoken.

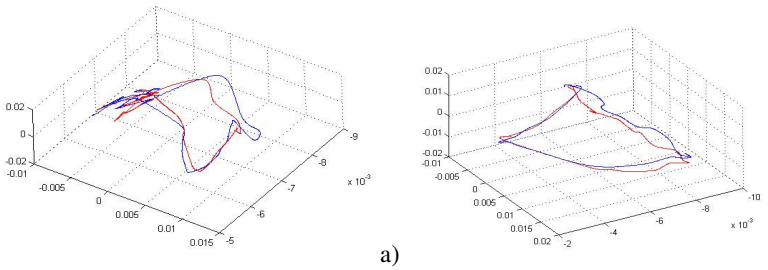
While the shape of the manifolds can be potentially used to discriminate between different words, they cannot be used directly to train a classifier or to recognize an unknown input image sequence. This is motivated by the fact that the number of frames contained in the input image is variable and depends on the complexity of the word spoken by the speaker. In this way, short words such as "I", "shy", etc. have associated a small number of frames and as results the manifolds will be defined by a small number of feature points. Conversely, longer words such as "banana" and "another" have associated larger image sequences and the number of feature points that defines the manifolds is larger. This is a real problem when these manifolds are used to train a classifier as the number of feature points is different. To circumvent this problem, we interpolate the feature points using a cubic spline and then re-sample uniformly the manifolds into a pre-defined number of keypoints. This procedure is appropriate as it allows a standard manifold re-sampling and they can be used to train a classifier or to obtain the classification result for an unknown image sequence.

3.3 Manifold interpolation using a cubic spline function

The application of cubic spline interpolation has two main advantages. Firstly, it allows us to generate a smooth surface for EM-PCA manifolds and secondly it reduces the effect of noise (and the influence of objects surrounding the lips such as teeth and tongue) associated with the feature points that form the manifolds in the EM-PCA space. This is clearly shown in Fig. 4 where we illustrate the appearance of the manifolds obtained after the application of cubic interpolation. Fig. 4 illustrates the interpolated manifolds generated for words "slow" and "shy".

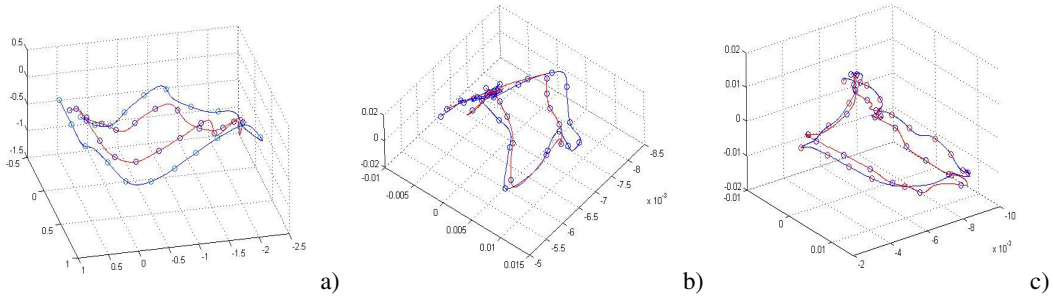
3.4 Manifold re-sampling into a predefined number of keypoints

As mentioned earlier, in order to generate standard data for training/recognition we need to uniformly re-sample the manifolds into a pre-defined number of keypoints. This re-sampling procedure will allow the identification of a standard set of keypoints as illustrated in Fig. 5. We decided to use uniform re-sampling since this procedure will generate keypoints that are equally distanced on the interpolated manifold surface and accurately sample the intrinsic information associated with the manifold's shape.



a) b)

Fig. 4. Manifolds resulting after cubic spline interpolation. (a) Word "Slow" – two image sequences; (b) Word "Shy" – two image sequences.



a) b) c)

Fig. 5. Uniform re-sampling of the interpolated manifolds (keypoints = 20) (a) two image sequences - word "Art"; (b) two image sequences – word "Slow"; (c) two image sequences - word "Shy". Note the good correspondence between the keypoints generated by re-sampling manifolds for similar words.

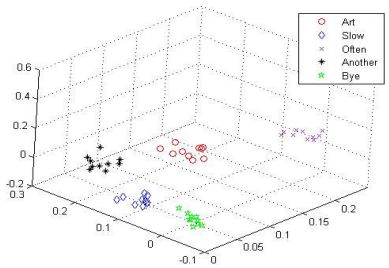


Fig. 6. Five words plotted in the second stage PCA space. Interpolated manifolds were re-sampled into 20 keypoints.

4 Classification

4.1 Second Stage EM-PCA and kNN classification

The second stage EM-PCA is applied to encode the temporal characteristics of the manifolds that are provided from the first stage of EM-PCA that is applied to generate the feature points that form the manifold [11]. For this implementation, this is achieved by taken the key points obtained after manifold re-sampling for each word

from the training data and compress this data using EM-PCA (only the largest three components were retained). In this way will be generated a data point for each class of words that are used to train a k-NN classifier (k=3). This is illustrated in Fig. 6 where 10 different instances of five words obtained after the application of the second stage PCA space are depicted. Once the training is complete the unknown datapoint is compressed using EM-PCA and is classified according to the distance to the nearest neighbour contained in the database [8].

4.2 HMM classifier

The second classification scheme evaluated in this paper is the HMM where the keypoints associated with each class of words are used directly for training [9, 16]. For this implementation we have constructed a HMM classifier for each word class, i.e. if we have *k* words in the database then we train *k* HMMs. The topology of the HMM classifier employed for this implementation is left to right and the number of states is set to three. The states of the HMM model are as follows. The first state encodes the transition from the initial state of the visual speech to articulation; the second state describes the articulation process while the last state models the transition from articulation to the end of the visual speech. These three states define the visual speech model that is used in the classification process. The length of sequence is set as the number of keypoints and maximum of number of iterations is set to 30.

5 Experimental Results

A number of experiments were carried out to assess the performance of the proposed system. For this study, we have created a database consisting of 10 words (30 examples for each word) generated by one speaker where 10 examples of each word are used for training and 20 examples are used for testing. The input database is divided into 10 classes and the classification results achieved by the kNN classifier when used in conjunction with second stage PCA are illustrated in Fig. 7. The classification results achieved by the HMM classifier are depicted in Fig. 8. By comparing the error recognition rate and the rate for incorrect recognition it can be concluded that the HMM classifier clearly outperforms the kNN classifier. Based on the evaluation of the experimental results we can observe that the best results are obtained when the interpolated manifolds are re-sampled to 20-30 keypoints. Another important finding resulting from this investigation is the fact that the manifolds offer a good discrimination (average classification success rate is 95% for HMM classifier) and they are suitable features to be used for visual speech recognition.

6. Conclusions

This paper describes the development of a visual speech recognition system where the main emphasis was placed on the evaluation of the discriminative power offered by a new manifold representation. In this regard, the manifolds are generated from the image data surrounding the lips and this data is compressed using an EM-PCA procedure into a low-dimensional feature space. Since these manifolds are defined by a different number of frames, they cannot be used directly as inputs for classification.

To address this problem we propose to interpolate the manifolds and then re-sample them uniformly into a predefined number of keypoints. It is useful to note that in our experiments, we have included image data that was generated by a single speaker and the word database contained a small number of words. We also noticed that the recognition rate of complexity words such as “banana” and “another” is generally lower than the recognition rate obtained for simpler words such as “I” and “shy”.

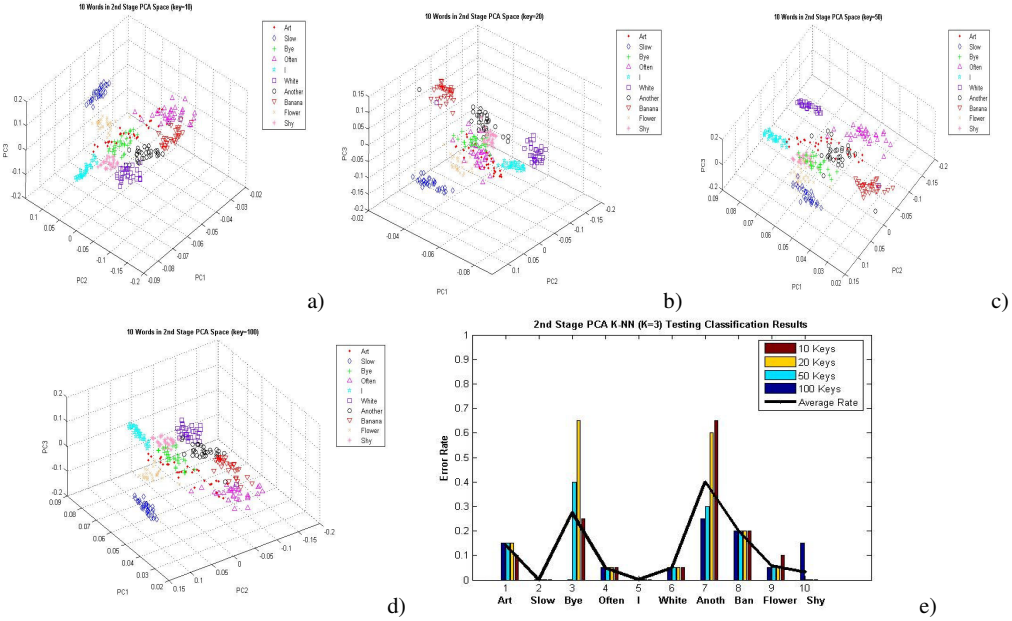


Fig. 7. Classification error rate achieved by the kNN classifier. (a) Manifold re-sampling: 10 keypoints; (b) Manifold re-sampling: 20 keypoints; (c) Manifold re-sampling: 50 keypoints; (d) Manifold re-sampling: 100 keypoints; (e) Classification accuracy (average 85%).

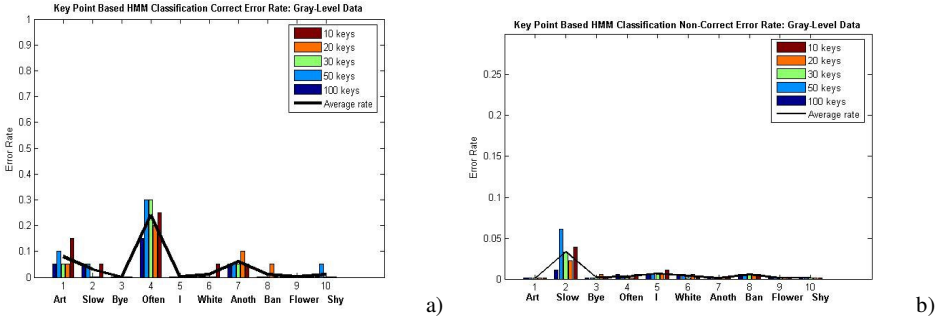


Fig. 8. Classification error rate achieved by HMM (a) Error recognition rate; (b) Rate for incorrect recognition. The average classification accuracy is 95%. From (a), it can be observed that the recognition rate is higher than that achieved by the kNN based classification scheme. We also can notice that the best result is obtained when the manifold re-sampling is set to 20 keypoints.

In the future, we aim to evaluate the proposed approach on a larger number of words are generated by different individuals and to include the temporal information as an additional cue in the recognition process.

References

1. N. Eveno, A. Caplier, P. Coulon, "Accurate and Quasi-Automatic Lip Tracking", IEEE Trans. Circuits Syst. Video Techn. 14(5), pp. 706-715, 2004.
2. Z. Ghahramani, Machine Learning Toolbox, Version 1.0 01-04-96, University of Toronto.
3. Y. L. Tian, T. Kanade (2000), "Robust lip tracking by combining shape colour and motion", Proc. of the Asian Conference on Computer Vision, pp. 1040 –1045, 2000.
4. A.V. Nefian, L.H. Liang, X. Liu, X. Pi, "Audio-Visual Speech Recognition", Intel Technology & Research, 2002.
5. N. Eveno, A. Caplier, and P.Y. Coulon. "A New Color Transformation for Lips Segmentation", IEEE Fourth Workshop on Multimedia Signal Processing, pp. 3-8, Cannes, France, 2001.
6. J. Luettin, N.A. Thacker, and S.W. Beet, "Active Shape Models for Visual Speech Feature Extraction", University of Sheffield, U.K., Tech. Rep. 95/44, 1995.
7. S. Roweis (1998), "EM Algorithms for PCA and SPCA", Advances in Neural Information Processing Systems, vol. 10, pp. 626-632.
8. T. Cootes, G. Edwards and C. Taylor, "A comparative evaluation of active appearance model algorithms" Proc of the British Machine Vision Conference, pp. 680-689, 1988.
9. S.W. Foo, Y. Lian, "Recognition of visual speech elements using adaptively boosted HMM", IEEE Trans. on Circuits and Systems for Video Technology, 14(5), pp. 693-705, 2004.
10. A. Shamaie and A. Sutherland, "Accurate Recognition of Large Number of Hand Gestures", Proc of Iranian Conference on Machine Vision and Image Processing, University of Technology, Tehran, 2003.
11. S.R. Das, R.C. Wilson, M.T. Lazarewicz, L.H. Finkel, "Gait recognition by two-stage principal component analysis", Automatic Face and Gesture Recognition, pp. 579-584, 2006
12. M. Gordan, C. Kotropoulos and I. Pitas, "Application of support vector machines classifiers to visual speech recognition", Proc. of the 2002 Int. Conf. on Image Processing, 2002.
13. X.P. Hong, H.X. Yao, Y.Q. Wan, R. Chen, "A PCA Based Visual DCT Feature Extraction Method for Lip-Reading", Proc. of Intelligent Information Hiding and Multimedia Signal Processing, pp. 321-326, 2006
14. W.C. Yau, D. K. Kumar, S. P. Arjunan, S. Kumar, "Visual Speech Recognition Using Image Moments and Multi-resolution Wavelet Images", Computer Graphics, Imaging and Visualisation, pp. 194-199, 2006
15. H.E. Cetingul, Y. Yemez, E. Erzin, A.M. Tekalp, "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading", IEEE Trans. on Image Processing, 15(10), pp. 2879-2891, 2006.
16. L. Dong, S.W. Foo and Y. Lian, "A two-channel training algorithm for Hidden Markov Model and its application to lip reading," EURASIP Journal on Applied Signal Processing, 2005(9), pp. 1382-1399, 2005.
17. R. Harvey, I. Matthews, J.A. Bangham, S. Cox, "Lip reading from scale-space measurements", Proc. of Computer Vision and Pattern Recognition, pp. 582-587, 1997.
18. C. Bregler, S.M. Omohundro,"Nonlinear manifold learning for visual speech recognition", Proc. of the International Conference on Computer Vision, pp. 494-499, 1995.