A New Manifold Representation for Visual Speech Recognition

Dahai Yu*, Ovidiu Ghita*, Alistair Sutherland, Paul F. Whelan*
Schools of Computing & Electronic Engineering, Vision Systems Group, Dublin City University*
Dahai.yu2@mail.dcu.ie

1. Introduction

In this paper, we propose a new manifold representation for visual speech recognition. The developed system consists of three main steps:

a. Lip extraction from input video data.
b. Generate the Expectation-Maximization PCA (EMPCA) manifolds for the entire image sequence and perform manifold interpolation and re-sampling.
c. Classify the manifolds using a HMM classifier to identify the words described by the lips motions in the input video sequence.

2. Manifold Representation

- Lips are segmented in each frame by thresholding the pseudo-hue component calculated from the RGB data.

- Encode each frame as a low-dimensional point in a feature space generated by the EM-PCA procedure.

- The low-dimensional feature points are joined by a polyline by ordering the frames in ascending order with respect to time. The surface defined by the polyline in the feature space is called manifold.

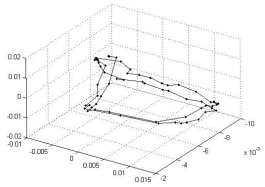
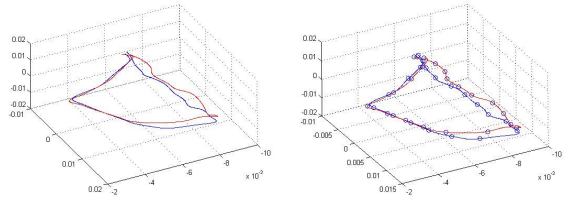


Fig. 1 Two manifolds associated with the same word spoken two times by the same speaker.

2.1. Manifolds Interpolation and Re-sampling

- To generate uniform data for classification the initial manifolds are first interpolated using a cubic spline and then re-sampled into a predefined number of key points (number of key points: 20, 50, 100).



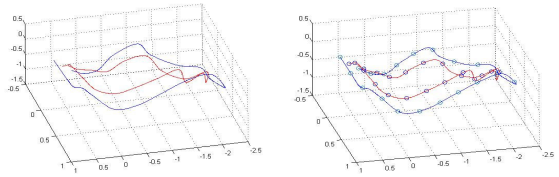


Fig. 2. Interpolated manifolds and the manifolds re-sampled into 20 uniformly distributed key points.

2.2. Advantages of Re-sampled Manifolds

- They encode in a compact model the shapes and the temporal distribution of the lips motions.

- The re-sampled manifolds offer robust word discrimination and can be easily used to train a HMM classifier and recognise the unknown words in an input video sequence.

3. Classification

- In this implementation we have employed the Hidden Markov Model (HMM) classification scheme.

- The observation data is defined by the re-sampled manifolds calculated from the training datasets.

4. Results

We have created a database consisting of 10 words with 30 examples of each word generated by one speaker. From these examples 10 were used for training and 20 examples were used for testing.

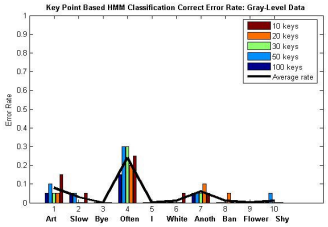


Fig. 3. Classification error rate.

5. Conclusions

The experiments indicate that the re-sampled manifolds are suitable features to be used for applications such as visual speech recognition and the average recognition rate is 95% when applied to the recognition of a set of 10 individual words.

References

[1] S.W. Foo, Y. Lian, "Recognition of visual speech elements using adaptively boosted HMM", IEEE Trans. on Circuits and Syst. for Video Tech., 14(5), pp. 693-705, 2004.

0-7695-2887-2/07 $25.00 © 2007 IEEE
DOI 10.1109/IMVIP.2007.35

IEEE computer society