A novel feature-level multiple HMMs classifier for Lipreading based on Ada-Boost Gabor kernels selection

Shengping Zhang, Hongxun Yao

School of Computer Science and Technology, Harbin Institute of Technology, China
{spzhang, yhx}@vilab.hit.edu.cn

Abstract

In this paper, a novel feature-level Multiple HMMs classifier for lipreading is presented. Firstly, it subdivides mouth images into four non-overlapping sub-blocks. Then AdaBoost is used to adaptively select optimal Gabor kernels from four sub-blocks convolved with different Gabor kernel functions and corresponding HMMs are trained. Finally the "boosted" HMMs are used to build a stronger multiple HMMs classifier by combining the decisions of the composite HMMs according to a probability synthesis rule. The method is evaluated on Bimodal Chinese Audio-Video Database (HIT Bi-CAVDB). Experimental results show that the proposed method gives distinctly superior recognition rate than traditional methods.

Keywords: Lipreading, AdaBoost, Gabor features, HMM

1. Introduction

Lipreading, as a multimodal human-computer intelligent interaction technology, has attracted increasing attention [3, 5, 6]. There are two important issues related to lipreading: 1) how to extract the most efficient features from lip sequences; 2) how to build a classifier for lipreading. This paper focuses on these two issues and presents a unified framework to address them.

Many feature extraction methods have been proposed in the literature over the past few decades. These methods extract features directly from mouth images after some image transforms such as Discrete Cosine Transform (DCT) [2], Principal Component Analysis (PCA) [3], Gabor Wavelets Transform (GWT) [4], etc. However problem one might encounter in these pixel-based methods is that the dimensionality of feature vector is extremely high, which would cause a "curse of dimensionality" problem [5]. For example, a 32×16 mouth image will get a 20480-dimensional feature vector after Gabor wavelets transform with 40 Gabor kernels in 5 scales and 8 directions. Such a high dimensional feature vector will cause two problems for lipreading. The first is that the degree of computational complexity is too high. It will restrict the applications of lipreading in real-time environment. The second is that the traditional Hidden Markov Model (HMM) cannot be used to model such high dimensional observation sequences.

The traditional HMM-based classifiers [1] [2] [3] only train a single HMM for a specific syllable with all the samples. There is a disadvantage that the final recognition result will be decided by the output of the single HMM classifier. However, the result of a single HMM would be wrong especially when the

training set is very small. In [6], an AdaBoost-HMM classifier was proposed, which trains multiple HMMs to cover different groups of training samples. However this sample level method has two shortcomings: 1) it only considers discriminative ability between vast samples, which are not available for realistic applications; 2) it ignores extracting discriminative features from a sample, which is more important for solving "curse of dimensionality" problem.

Inspired by the work in [6], we present a novel feature-level multiple HMMs classifier for lipreading. The approach utilizes the appearance symmetry of the mouth image and the directions of Gabor kernels. Firstly, it subdivides mouth images into four non-overlapping sub-blocks. Then AdaBoost is used to adaptively select optimal Gabor kernels from four sub-blocks convolved with different Gabor kernel functions and corresponding HMMs are trained using biased Baum-Welch method. Finally the "boosted" HMMs are used to build a stronger multiple HMMs classifier by combining the decisions of the composite HMMs according to a probability synthesis rule.

The remainder of this paper is laid out as follows. The block-based Gabor wavelets transform is presented in Section 2. In Section 3, we describe the multiple-HMM classifier using Adaboost technique. Experiments and analysis are shown in section 4, followed by some conclusion in section 5.

2. Features extraction using block-based Gabor wavelets transform

2.1. Gabor kernel functions

Chui [7] gives a good introduction to image representation using Gabor functions. A Gabor kernel function is the product of an elliptical Gaussian envelope and a complex plane wave, defined as:

$$\psi_{s,d}(x,y) = \psi_{\bar{k}}(z) = \frac{\|\bar{k}\|}{\delta^2} \cdot e^{\frac{\|\bar{k}\|\|z\|^2}{2\delta^2}} \cdot \left[e^{i\bar{k}\cdot z} - e^{-\frac{\delta^2}{2}} \right] \quad (1)$$

where $z = (x, y)$ is the coordinates of pixel and $\bar{k}$ is the frequency vector, which determines the scale and direction of Gabor functions, $\bar{k} = k_s e^{i\phi_d}$, where $k_s = k_{\max} / f^s$, $k_{\max} = \pi/2$. In ordinary application, $f = 2$, $s = 0,1,2,3,4$, and $\phi_d = \pi d/8$, for $d = 0,1,2,3,4,5,6,7$. Examples of the real part of Gabor kernel functions are presented in Fig. 1.

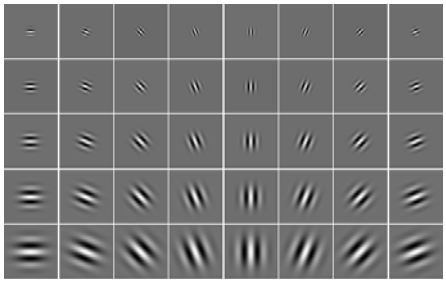

Fig. 1: The real part of Gabor kernel functions with 5 scales and 8 directions.

Fig. 2: The entire mouth image is divided into 4 non-overlapping sub-blocks.

2.2. Block-based Gabor wavelets transform

The Gabor transform results of an image are obtained by convolving the Gabor kernels (1) with the image as follows:

$$o_{s,d}(x,y) = I(x,y) * \psi_{s,d}(x,y) \quad (2)$$

where $s = 0,1,2,3,4$ and $d = 0,1,2,3,4,5,6,7$. There are 40 com-

ponents, and each one is the magnitude part of the output (2). A 32×16 mouth image will get a 20480-dimensional feature vector after convolved with 40 Gabor kernels. The dimensionality of resulting vector is much larger than the number of the samples in the training set, which leads to the "curse of dimensionality" problem. In consideration of the appearance symmetry of mouth region and the directions of Gabor kernels, we first subdivide the 32×16 mouth region into 4 non-overlapping sub-blocks of size 16 × 8 as shown in Fig. 2, and then we convolve each sub-block with 40 Gabor kernels. The block-based Gabor wavelets transform results are shown in Fig. 3.

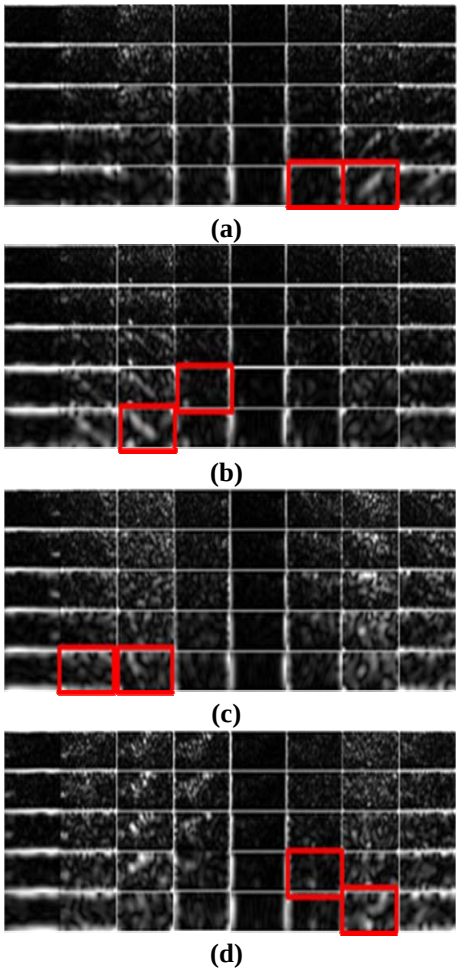

Fig. 3: (a), (b), (c) and (d) are block-based Gabor wavelets results on upper-left, upper-right, lower-left and lower-right sub-blocks respectively. The blocks marked in red are the boosted sub-vector by proposed method.

The resulting feature vector obtained by block-based Gabor wavelets transform consists of 160 sub-vectors as follows:

$$O = \{o_{b,s,d}\} \tag{3}$$

where $b = 0,1,2,3$, $s = 0,1,2,3,4$, $o = 0,1,2,3,4,5,6,7$, b is the index of sub-blocks. Its values 0, 1, 2, 3 denote the upper-left, upper-right, lower-left and lower-right blocks, respectively. As shown in Fig. 3, we can find that the discriminative ability of every sub –block convolved with different direction Gabor kernels is different.

3. Feature-level multiple HMMs classifier based on AdaBoost Gabor kernels selection

In a traditional single-HMM classifier of identifying K syllable, K HMMs $\theta_1, \theta_2, ..., \theta_K$ are included. The input to the system is the observation sequences of an unknown syllable, denoted by x^T . In the training stage, each HMM is trained with the samples of a specific syllable class. Each HMM determines the probability $P\left(x^T \mid \theta_k\right)$ of occurrence of the unknown syllable. The final decision is made by comparing the probability $P\left(x^T \mid \theta_k\right)$ for all the classes. The one that give the maximum probability is chosen as the identity of x^T , as follows:

$$ID(x^T) = \arg\max_k P\left(x^T \mid \theta_k\right), \quad (k = 1,2,...,K) \tag{4}$$

Inspired by the work in [8], in order to address the "curse of dimensionality" of Gabor features, we present a novel feature-level multiple HMMs classifier

based on AdaBoost Gabor kernels selection. Initially, we put all the 160 sub-vectors (3) into a feature pool. At each iteration, certain weights are assigned to the remainder sub-vectors in feature pool. We use the biased Baum-Welch method [6] to train the HMMs for each sub-vector in feature pool. Resulting HMMs corresponding to a sub-vector are then used to recognize the training samples and get an error rate. We choose a sub-vector which has the minimum error rate and remove it from the feature pool. The corresponding HMMs will be chosen as composite HMMs of the final multiple-HMMs classifier. The algorithm is as follow:

Input: Training set $(x_1, y_1),...,(x_n, y_n)$,

$y_i = 0,1,...K$ for K syllables class

1. Start with weights distribution $D_1(x_i) = 1/n$. The feature pool consists of 160 sub-vectors for each sample.
2. Repeat for t = 1, …, T

1) For each sub-vector $o_{b,s,d}$ in feature pool, train K new HMMs $\theta_k^t \ (k = 1,2,...,K)$ using the biased Baum-Welch algorithm. The error is evaluated with respect to weights distribution D_t ,

$$\varepsilon_{b,s,d} = \sum_i D_t(x_i) \cdot s\left(ID(x_i) - y_i\right)$$

2) Choose the sub-vector $o_{b,s,d}$ with the lowest error ε_t , remove it from feature pool. Corresponding HMMs are chosen as a composite HMMs $\Theta_k^t \ (k = 1,2,,,K)$

3) Update the weights: $D_{t+1}(x_i) = D_t(x_i)\beta_t^{1-e_i}$ wher e $e_i = 0$ if sample x_i is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$

4) Normalize the weights:

$$D_{t+1}(x_i) \leftarrow \frac{D_t(i)}{\sum_{j=1}^{n} D_t(j)}$$

Output: The final multiple-HMMs classifier:

$$ID(x^T) = \arg\max_k \sum_{t=1}^{T} \log\left(\frac{1}{\beta_t}\right) P\left(x^T \mid \Theta_k^t\right) \quad (5)$$

where $s(x) = \begin{cases} 0, x = 0 \\ 1, x \neq 0 \end{cases}$ and Θ_k^t is the boosted HMM for kth class at tth iteration. T is a user-settable parameter, which specifies the number of sub-vectors needed to be selected. The procedure terminates when T iterations are completed.

4. Experimental results and analysis

In order to confirm the effectiveness of the proposed method, we conduct experiments on the Harbin Institute of Technology Bimodal Chinese Audio-Video Database (HIT Bi-CAVDB) [4].

Considering the dimensionality of feature vector and computational complexity, we set T to be eight. The eight sub-vectors selected by proposed method are $o_{0,4,6}$, $o_{0,4,5}$, $o_{1,3,3}$, $o_{1,4,2}$, $o_{2,4,1}$, $o_{2,4,2}$, $o_{3,4,6}$ and $o_{3,3,5}$. We marked these sub-vectors in Fig. 3. As shown in Fig. 3, we can find that the selected sub-vectors for each sub-block have the almost consistent directions with the appearance of the sub-block image. The Gabor kernels with horizontal and vertical directions at any scale have lower discriminative ability. We get the recognition rate with T increasing from 1 to 8. As shown in Fig. 4, the recognition rate dramatically increases as T increases from 1 to 5. However, the increase slows down as T increases from 6 to 8. Our best recognition rate reaches 83.7% with T set to be 8. It can be explained that the extracted features of four sub-blocks sup-

plement each other when T is small. However, as T keeps on increasing, the recognition rate will trend to stabilization because of the feature redundancy caused by their appearance symmetry.

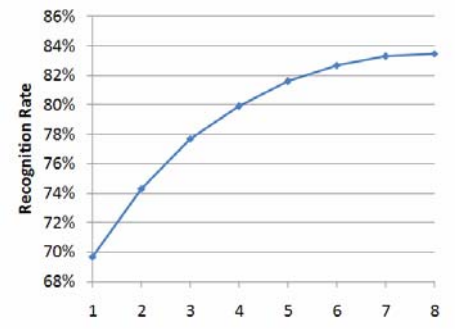

Fig. 4: The recognition rates with T increasing from 1 to 8

We also compare the proposed multiple HMMs method with two traditional single-HMM methods: DCT based and FFT based methods, which extract features using DCT and FFT, respectively. The feature dimensionalities of three methods are reduced to 60 by principal component analysis. The comparative results are shown in Fig. 5 and demonstrate that the proposed method gives distinctly superior recognition rate than traditional methods.

A: DCT Based method B: FFT Based method C: Proposed method

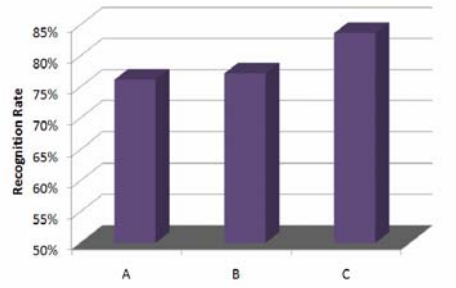

Fig. 5: Comparison results with traditional methods

5. Conclusion

This paper presents a novel feature level multiple-HMMs classifier method for lipreading, which utilizes the appearance symmetry of mouth image and the directions of Gabor kernels. It first subdivides mouth images into four non-overlapping sub-blocks and then uses AdaBoost to adaptively select optimal Gabor kernels on feature level for every sub-block. Proposed method integrates the feature selection and classifier building into a unified framework, which effectively addresses the "curse of dimensionality" problem and overcome the shortcomings of single-HMM classifier. Experimental results show that proposed method outperforms traditional methods.

6. Acknowledgement

This research is supported by the Program for New Century Excellent Talents in University (NCET -05-03 34) and the Natural Science Foundation of Heilongjiang province (Grant No. E2005-29).

7. References

[1] E. D. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D. dissertation, Univ. Illinois, Urbana-Champaign, 1984.

[2] G.Potamianos, H.P. Graf, and E.Cosatto. "An image transform approach for HMM based automatic lipreading." Proc. Int. Conf. Image Process, Chicago, pp.173-177, 1998.

[3] S.Dupont and J.Luettin, "Audio-visual speech modeling for continuous speech recognition", IEEE Trans. On Multimedia, 2:141-151,2000.

[4] S. P. Zhang, H. X. Yao, Y. Y. Wan, and D. Wang. "Combining Global and Local Classifiers for Lipreading". ACII 2007: 733-734.

[5] Matthews, etc. "Extraction of Visual Features for Lipreading". IEEE Trans. on Pattern Analysis and Machine In-

telligence, Vol. 24, No. 2, February 2002.
[6] S. W. Foo, Y. Lian, and L. Dong. "Recognition of Visulal Speech Elements Using Adaptively Boosted Hidden Markov Models", IEEE Trans. On Circuits and systems for video technology, VOL. 14, 693-705, 2004.
[7] C. K. Chui, An Introduction to Wavelets. Boston, MA: Academic,1992.
[8] R. E. Schapire, "A brief introduction to boosting," in 16th Int. Joint Conf. Artificial Intelligence, 1999, pp. 1305-1401.