Combining Global and Local Classifiers for Lipreading

Shengping Zhang, Hongxun Yao, Yuqi Wan, and Dan Wang

School of Computer Science and Engineering, Harbin Institute of Technology,
Harbin, 150001, China

Lipreading has become a hot research topic in recent years since the visual information extracted from the lip movement has been shown to improve the performance of automatic speech recognition (ASR) system especially under noisy environments [1]-[3], [5]. There are two important issues related to lipreading: 1) how to extract the most efficient features from lip image sequences, 2) how to build lipreading models. This paper mainly focuses on how to choose more efficient features for lipreading.

Feature extraction is very important for lipreading. Many feature extraction methods have been proposed in the literature. In general, variant feature extraction methods can be divided into two kinds: 1) pixel-based features derived directly from the image transforming such as Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) [2] [6], Principal Component Analysis (PCA) [3], Linear Discriminant Analysis (LDA) [6], Gabor Wavelets Transform(GWT) [4], Local Binary Pattern (LBP)and so on. 2) model-based features by tracking lip contours to describe its movement [5]. Some experiment results show that pixel-based method is more efficient than model-based method [2] [5].

In those pixel-based feature extraction methods mentioned above, DFT and DCT extract the global features in the mouth images; GWT and LBP extract local features. Global features consider the mouth image as a whole and it is easy to reflect the whole difference of mouth images. Local features, on the other hand, are computed at multiple points in the mouth images and are more robust to the variations between the images of the same mouth due to illumination and viewing direction. Although both global and local features work well to some extent, each is limited by the fact that it ignores other information that may also be very important.

Most lipreading systems tent to use either global or local features. Some psychological evidences show that people use both global and local features for object recognition, in some extent, people use global features before analyzing the image in detail [7] [8]. Local features could result in much better performance than global ones. Motivated by this study, this paper presents a novel method of combining global and local classifiers to form a more powerful classifier for lipreading. The global classifier uses Discrete Fourier Transform (DFT) to extract global features. The local classifier uses block-based Gabor Wavelets Transform (BGWT) to extract local features. Both global and local classifiers use Hidden Markov Models (HMM) to model. These two classifiers are then

A. Paiva, R. Prada, and R.W. Picard (Eds.): ACII 2007, LNCS 4738, pp. 733–734, 2007.

combined to form the final classifier which not only uses the global information but also the local information.

We investigate and compare several current popular global and local feature extraction methods such as Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), block-based Gabor wavelets Transform (BGWT) and Local Binary Pattern (LBP) as well as several combination methods between them. The experiment results reveal that the performance of single global or local classifier is around 77%. Among these classifiers, the one based on LBP performs worse than others. The reason is that LBP is implemented in spatial domain and extracted global features are not more efficient than other features which are extracted in frequency domain. In all of the combinations of global and local classifiers, the combination of DFT classifier and Gabor classifier (DFT+BGWT) gained the highest accuracy up to 82.45% which not only surpasses each of the individual classifiers but also the other combinations, in other words, the global DFT features can compensate the local Gabor features much better. The combination of DCT classifier and Gabor classifier gains worse recognition rate than DFT+BGWT. In fact, DCT coefficients can be derived from the real part of DFT coefficients, and DFT features have more powerful capability to reflect the intensity variations in an entire image. The performance of combination of DFT and LBP is better than LBP but worse than DFT. The reason is that LBP features have worse discriminability than DFT features. When they are used together, the test samples which will be recognized right by DFT may be recognized wrong with the influence of weaker LBP classifier.

References

1. Morishima, S., Ogata, S., Murai, K., Nakamura, S.: Audio-visual speech translation with automatic lip synchronization and face tracking based on 3D head model. Proc. IEEE Int. Conf. Acoustics, Speech,and Signal Processing 2, 2117–2120 (2002)
2. Potamianos, G., Graf, H.P., Cosatto, E.: An image transform approach for HMM based automatic lipreading. In: Proc. Int. Conf. Image Process, Chicago, pp. 173–177 (1998)
3. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. On Multimedia 2, 141–151 (2000)
4. Shen, L., Bai, L.: Gabor feature based face recognition using kernel methods. AFGR, pp. 170–176 (2004)
5. Matthews., et al.: Extraction of Visual Features for Lipreading. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(2) (2002)
6. Duchnowski, P., et al.: Toward movement-invariant automatic lip-reading and speech recognition. In: Duchnowski, P. (ed.) Proc. Int. Conf. Acoust. Speech Signal Process, Detroit, pp. 109–111 (1995)
7. Navon, D.: Forest before the trees: the precedence of global features in visual perception. Cognitive Psychology 9, 353–383 (1977)
8. Biederman, I.: On the semantics of a glance at a scene. In: Kubovy, M., Pomerantz, J. (eds.) Perceptual organization, pp. 213–253. Erlbaum (1981)