Audio and Video Reactive Talking Head Avatar

S. Pachoud[a], S. Gong[a], A. Cavallaro[b]

[a]Department of Computer Science
Queen Mary, University of London
{spachoud,sgg}@dcs.qmul.ac.uk

[b]Department of Electronic Engineering
Queen Mary, University of London
andrea.cavallaro@elec.qmul.ac.uk

Human perception is multi-sensory. In particular, one often uses two of our five senses: Sight and Hearing. Sight or vision describes the ability to detect electromagnetic waves within the visible range (light) by the eye and the brain to interpret the image as sight. Hearing or audition is the sense of sound perception and results from tiny hair fibres in the inner ear detecting the motion of a membrane. For perceiving facial emotion and behaviour, humans combine the acoustic waveform (audio information) and the movements of the lips, tongue and other facial muscles (visual information) generated by a speaker. The McGurk effect [1] established this bimodal speech perception by showing that, when conflicting audio and visual stimuli is produced to an individual, the latter may assimilate a new stimulus, different from the two others. For example if you watch a talking head, in which repeated utterances of the syllable [ba] had been dubbed on to lip movements for [ga], normal adults reported hearing [da]. Effective human computer interaction (HCI) requires a multimodal speech recognition in order to make the computer interplays with human in the same way as human to human communication.

Such observations have motivated interest in automatic recognition of visual speech. The researches in this field aim to improve the speech recognition systems by taking advantage of the visual modalities of a speaker in addition to the usual audio modalities. The purpose of this combination is to ensure that the resulting system performance is more accurate than that of each single modality alone. There are three main areas that compose audio-visual automatic speech recognition (AV-ASR) systems: The front end design, the audio-visual integration strategy, and the speech recognition method used. Fig 1 shows this general architecture, divided into two different integration schemes: on the left, feature fusion class and on the right decision fusion class. With the first class, the same recognition algorithms are used to concatenate audio and visual features. In the second one, separate recognizers are trained for the A+V features. Audio-visual fusion can also occur at a level between feature and decision classes.

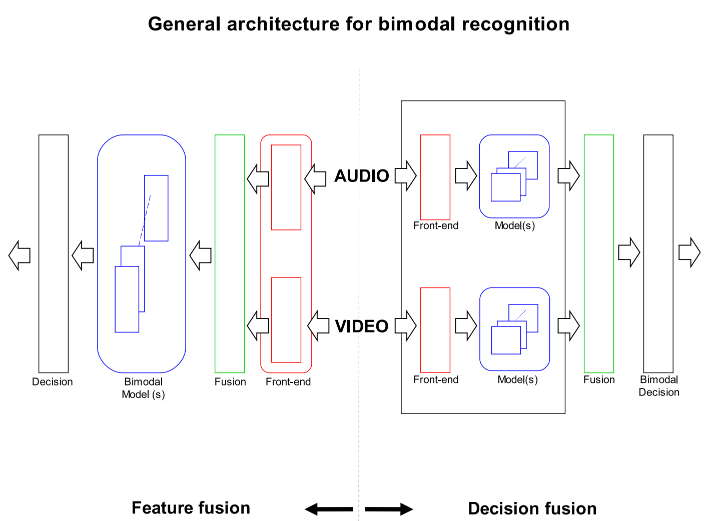

Figure 1: General architecture for bimodal recognition.

The Audio Visual Speech Recognition group of IBM® published several publications [2, 3, 4] on AV-ASR. Their bottom-up, or texture-based, system consists of a two-dimensional discrete cosine transform (DCT), applied to the lip image. To reduce dimensionality and improve discrimination among the speech classes, an intra-frame linear discriminant analysis (LDA) projection is applied. However, lip texture features are sensitive to intensity variations between the training and test data sets. Wang et al. [5] and Matthews et al. [6] developed a top-down, or shape-based, approach to extract lip features. It requires some prior assumptions of what are important visual speech features. With this approach, it is easier to match mouth openings-closings with corresponding phonemes. However it suffers from complex feature extraction and training processes. Moreover only isolated English digits (0-9) or letters (A-Z) are used in their experiments. A third alternative representation for lip information is used in [7]: an explicit lip motion information. The authors claim that a grid-based dense lip motion features are superior and more robust compared to shape-based lip motion features. Nevertheless, their lip contour extraction needs manual manoeuvres, which are incompatible with real-world application. Recently, Göcke [8] proposed a non-intrusive[1] real-time 3D lip tracking algorithm. His experiments are only performed in an ideal environment, without any visual and audio noise.

[1]No facial markers or special make-up.

More specifically, the general approach for lip reading, reported in the literature, is to extract the principal components of the lip movement in order to establish a one-to-one correspondence between phonemes of speech and visemes of lip shape [2, 3, 5]. Alternatively, we consider to compare the movements of lips generated by a speaker (called query movements) with the movements of lips of all visemes in a certain language (called reference movements). The system has to be able to recognise the original visemes of the database (reference movements) apart from the person (gender, age, language,...). Our approach aims to define the features in space and over time.
To confirm the reliability of working in a spatio-temporal space, we first work in static spacial feature alone: we try to find similarities between frames (or pictures). We start with Boiman and Irani's model [9]. They developed an approach for measuring similarity between two signals. They derive principled information-theoretic measures for local and global similarity between a query (one or more signals) and a reference (one or more signals). In our first purpose (find similarities between frames), the query or the reference signal is an image. Our approach works as follow: At first, we divide the query image in small macro-block (MB_q^i with ($i = 1, 2, \ldots, |\mathrm{MB}_q|$)). Then each MB_q^i is divided in nine blocks (B_q^j with ($i = 1, 2, \ldots, 9$)). For those two division, there is no overlapping. Then we look for similarities between each MB_q^i and the reference image. To perform this, we scan each B_q^j of each MB_q^i in the reference image and the probability of a macro-block to be in the reference image is the following:

$$P(\mathrm{MB}_q^i, ref) = \prod_i e^{-\frac{|\Delta d_i|^2}{2\sigma_d^2}} e^{-\frac{|\Delta l_i|^2}{2\sigma_l^2}} \quad (1)$$

where Δd_i and Δl_i are the differences between and the local displacements of the descriptors of B_q^j and its correspondent in the reference image respectively. σ_d and σ_l are the only parameters we have to define. σ_d is determined empirically and σ_l is equal to the norm of the diagonal of a macro-block. Our descriptors are calculated using SIFT [10]. Fig. 2 shows a graph and an explanation of the process we have made until now. Picture (a) is the query image. Then we create small macro-blocks (MB_q). In (b), we see one of them in green and for the next pictures ((c) to (e)), the development is done with only one macro-block. Then, we extract each blocks (c) and we divide them in nine part as we see in (d) in blue. Next, we look for each MB_q separately in the reference image (e). We perform this without overlapping during the scanning of the MB_q as we show in yellow in (f). In (g), a bar graph of probabilities $P(\mathrm{MB}_q^i, ref)$ (see eq. 1) for each macro-blocks is drawn. The three highest probability of a MB_q to be in the reference image are showed in red in (h).

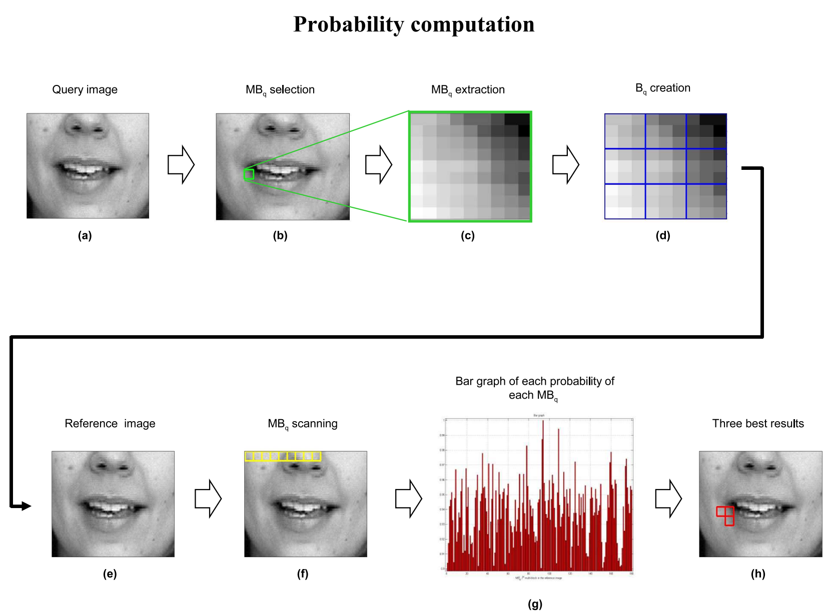

Figure 2: Graph of our process.

First experimental results show that there is not enough discrimination between each $P(\mathrm{MB}_q^i, ref)$ for one MB_q. That is due because of the insufficiency of either our features or our measure (or both). We currently consider the use of additional features such as Haar wavelets. Moreover the spacial feature does not give enough information for out task, so we are extended our model to spatio-temporal space.

References

[1] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, December 1976.

[2] J. Huang, G. Potamianos, and C. Neti. Improving audio-visual speech recognition with an infrared headset. In *Audio-Visual Speech Processing, 2003. International Conference on*, pages 175–178, St-Jorioz, France, Semptember 2003.

[3] G. Potamianos, C. Neti, J. Huang, J.H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas, and J. Jiang. Towards practical deployment of audio-visual speech recognition. In *Acoustics, Speech, and Signal Processing, 2004. IEEE International Conference on*, volume 3, pages iii–777–780, 2004.

[4] J. Huang, E. Marcheret, and K. Visweswariah. Rapid feature space speaker adaptation for multi-stream hmm-based audio-visual speech recognition. In *Multimedia and Expo, 2005. IEEE International Conference on*, pages 338–341, 2005.

[5] S.L. Wang, W.H. Lau, S.H. Leung, and H. Yan. A real-time automatic lipreading system. In *Circuits and Systems, 2004. International Symposium on*, volume 2, pages II–101–4, 2004.

[6] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *Pattern Analysis and Machine Intelligence. IEEE Transactions on*, 24(2):198–213, 2002.

[7] H. Ertan Çetingül, Yücel Yemez, Engin Erzin, and A. Murat Tekalp. Discriminative analysis of lip motion features for speaker identification and speech-reading. *Image Processing. IEEE Transactions on*, 15(10):2879–2891, 2006.

[8] Roland Göcke. Audio-video automatic speech recognition: an example of improved performance through multimodal sensor input. In *NICTA-HCSNet Multimodal User Interaction Workshop, 2005*, volume 57, pages 25–32, Darlinghurst, Australia, 2006.

[9] O. Boiman and M. Irani. Similarity by composition. In *Neural Information Processing Systems*, Vancouver, December 2006.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2004.