Chapter XIII

Motion Features for Visual Speech Recognition

Wai Chee Yau
RMIT University, Australia

Dinesh Kant Kumar
RMIT University, Australia

Hans Weghorn
BA University of Cooperative Education Stuttgart, Germany

ABSTRACT

The performance of a visual speech recognition technique is greatly influenced by the choice of visual speech features. Speech information in the visual domain can be generally categorized into static (mouth appearance) and motion (mouth movement) features. This chapter reviews a number of computer-based lip-reading approaches using motion features. The motion-based visual speech recognition techniques can be broadly categorized into two types of algorithms: optical-flow and image subtraction. Image subtraction techniques have been demonstrated to outperform optical-flow based methods in lip-reading. The problem with image subtraction-based method using difference of frames (DOF) is that these features capture the changes in the images over time, but do not indicate the direction of the mouth movement. New motion features to overcome the limitation of the conventional image subtraction-based techniques in visual speech recognition are presented in this chapter. The proposed approach extracts features by applying motion segmentation on image sequences. Video data are represented in a 2-D space using grayscale images named as motion history images (MHI). MHIs are spatio-temporal templates that implicitly encode the temporal component of mouth movement. Zernike moments are computed from MHIs as image descriptors and classified using support vector machines (SVMs). Experimental results demonstrate that the proposed technique yield a high accuracy in a phoneme classification task. The results suggest that dynamic information is important for visual speech recognition.
INTRODUCTION

Speech recognition technologies provide the flexibility for users to control computer through speech. The difficulty of speech recognition systems based on acoustic signals is the sensitivity of such systems to variations in acoustic conditions. The performance of audio speech recognizers degrades drastically when the acoustic signal strength is low, or in situations with high ambient noise levels. To overcome this limitation, there is an increasing trend in applying non-acoustic modalities in speech recognition. A number of alternatives have been proposed, such as visual (Petajan, 1984), recording of vocal cords movements through electroglottograph (EGG) (Dikshit, 1995) and recording of facial muscle activity (Arjunan, 2007). Vision-based techniques are non intrusive and do not require the placement of sensors on a speaker’s face and hence are the more desirable options.

The use of visual signals in computer speech recognition is consistent with the way human perceive speech. Human speech perception consists of audio and visual modalities which are demonstrated by McGurk effect. McGurk effect occurs in situations when normal hearing adults are presented with conflicting visual and audio speech signals, the perception of sound is changed (McGurk & MacDonald, 1976). An example is when a listener hears a sound of /ba/ and sees a lip movement of /ga/, the sound /da/ is perceived. This indicates that substantial amount of speech information is encoded in visual signals. Visual speech information has been demonstrated to improve robustness of audio-only speech recognition systems (Stork & Hennecke 1996; Potamianos, Neti, Gravier, Garg, & Senior, 2004; Aleksic & Katsaggelos, 2005).

The visual cues contain far less classification power for speech as compared to audio data and hence it is to be expected that the visual-only speech recognition would support only a small vocabulary. High accuracies are achievable for small vocabulary, speaker-dependent visual-only speech recognition problems as reported in (Nefian, Liang, Pi, Liu & Murphy, 2002; Zhang, Mersereau, Clements & Broun, 2002; Foo & Dong, 2002). An increase in the number of speakers and size of the vocabulary would result in degradation of the accuracy of visual speech recognition. This is demonstrated by the high error rates reported by Potamianos et al. (2003) and Hazen (2006) in large vocabulary visual-only speech recognition task, with errors of the order of 90%. Further, these errors are also attributed to the large inter-subject variations caused by the differences in lip movements for the same utterance spoken by different speakers. The use of visual speech information for speaker recognition (Luettin, Thacker & Beet, 1996; Faraj & Bigun 2007) indicates the large variations that exit between the speaking styles of different people. This difference is even greater if we transgress across the geographic and cultural boundaries.

A typical visual speech recognition technique consists of three phases, (i) recording and preprocessing of video data, (ii) extraction of visual speech features and (iii) classification. One of the main challenges in visual speech recognition is the selection of features to represent lip dynamics. Visual speech features contain information on the visible movement of speech articulators such as lips, teeth and jaw. Various visual speech features have been proposed in the literature. These features can be broadly categorized into shape-based (model-based), appearance-based and motion features. Shape-based features rely on the geometric shape of lips.

Visual speech recognition was first proposed by Petajan (1984) using shape-based features such as height and width of the mouth. Researchers have reported the use of artificial markers on user’s mouth to extract lip contours (Kaynak, Qi, Cheok, Sengupta, & Chung, 2001). The use of artificial markers
Related Content

**Bangla Character Recognition Using Optical Joint Transform Correlation**
[www.igi-global.com/chapter/bangla-character-recognition-using-optical/78473?camid=4v1a](www.igi-global.com/chapter/bangla-character-recognition-using-optical/78473?camid=4v1a)

**Tools for the Automatic Generation of Ontology Documentation: A Task-Based Evaluation**
[www.igi-global.com/chapter/tools-for-the-automatic-generation-of-ontology-documentation/108754?camid=4v1a](www.igi-global.com/chapter/tools-for-the-automatic-generation-of-ontology-documentation/108754?camid=4v1a)

**Perception of Vowels and Dental Consonants in Bangla Speech Processing**
[www.igi-global.com/chapter/perception-vowels-dental-consonants-bangla/78478?camid=4v1a](www.igi-global.com/chapter/perception-vowels-dental-consonants-bangla/78478?camid=4v1a)

**Visual Speech and Gesture Coding Using the MPEG-4 Face and Body Animation Standard**
[www.igi-global.com/chapter/visual-speech-gesture-coding-using/31066?camid=4v1a](www.igi-global.com/chapter/visual-speech-gesture-coding-using/31066?camid=4v1a)