Visual Speech Recognition:
Lip Segmentation and Mapping

Alan Wee-Chung Liew
Griffith University, Australia

Shilin Wang
Shanghai Jiaotong University, China

Medical Information Science REFERENCE

MEDICAL INFORMATION SCIENCE REFERENCE

Hershey · New York

Table of Contents

Section II
Lip Modeling, Segmentation, and Feature Extraction

Section III
Visual Speech Recognition

Detailed Table of Contents

There has been significant work on investigating the relationship between articulatory movements and vocal tract shape and speech acoustics (Fant, 1960; Flanagan, 1965; Narayanan & Alwan, 2000; Schroeter & Sondhi, 1994). It has been shown that there exists a strong correlation between face motion, and vocal tract shape and speech acoustics (Grant & Braida, 1991; Massaro & Stork, 1998; Summerfield, 1979, 1987, 1992; Williams & Katsaggelos, 2002; Yehia, Rubin, & Vatikiotis-Bateson, 1998). In particular, dynamic lip information conveys not only correlated but also complimentary information to the acoustic speech information. Its integration into an automatic speech recognition (ASR) system, resulting in an audio-visual (AV) system, can potentially increase the system's performance. Although visual speech information is usually used together with acoustic information, there are applications where visual-only (V-only) ASR systems can be employed achieving high recognition rates. Such include small vocabulary ASR (digits, small number of commands, etc.) and ASR in the presence of adverse acoustic conditions. The choice and accurate extraction of visual features strongly affect the performance of AV and V-only ASR systems. The establishment of lip features for speech recognition is a relatively new research topic. Although a number of approaches can be used for extracting and representing visual lip information, unfortunately, limited work exists in the literature in comparing the relative performance of different features. In this chapter, we will describe various approaches for extracting and representing important visual features, review existing systems, evaluate their relative performance in terms of speech and speaker recognition rates, and discuss future research and development directions in this area.

Chapter III

Lip segmentation is the first step of any audio-visual speech reading system. The accuracy of this segmentation has a major influence on the performances of the global system. But this is a very difficult task. First of all, lip shape can undergo strong deformations during a speech sequence. As many other image processing algorithms, the segmentation task is also influenced by the illumination conditions and by the orientation of the object to be segmented. In this chapter, we present an overview about lip modeling and lip segmentation (region-based and contour-based methods). We limit our study to the problem of lip segmentation in frontal faces. Section 1 gives an overview about the chrominance information that is used for lip segmentation and a comparison between different chrominance cues is proposed. Section 2 presents region-based approaches and training steps. Section 3 focuses on contour-based approaches and parametric lip models. Section 4 inventories methods for lip segmentation accuracy evaluation. Some specific applications are briefly presented in section 5.

Chapter IV

Automatic Speech Recognition (ASR) is the most natural input modality from humans to machines. When the hands are busy or a full keyboard is not available speech input is especially in demand. Since

the most compelling application scenarios for ASR include noisy environments (mobile phones, public kiosks, cars), visual speech processing must be incorporated to provide robust performance. This chapter motivates and describes the MPEG-4 Face and Body Animation (FBA) standard for representing visual speech data as part of a whole virtual human specification. The super low bit-rate FBA codec included with the standard enables thin clients to access processing and communication services over any network including enhanced visual communication, animated entertainment, man-machine dialog, and audio/visual speech recognition.

Section II
Lip Modeling, Segmentation, and Feature Extraction

As the first step of many visual speech recognition and visual speaker authentication systems, robust and accurate lip region segmentation is of vital importance for lip image analysis. However, most of the current techniques break down when dealing with lip images with complex and inhomogeneous background region such as mustaches and beards. In order to solve this problem, a Multi-class, Shape-guided FCM (MS-FCM) clustering algorithm is proposed in this chapter. In the proposed approach, one cluster is set for the lip region and a combination of multiple clusters for the background which generally includes the skin region, lip shadow or beards. With the spatial distribution of the lip cluster, a spatial penalty term considering the spatial location information is introduced and incorporated into the objective function such that pixels having similar color but located in different regions can be differentiated. Experimental results show that the proposed algorithm provides accurate lip-background partition even for the images with complex background features.

An algorithm for lip contour extraction is presented in this chapter. A colour video sequence of a speaker's face is acquired under natural lighting conditions without any particular set-up, make-up, or markers. The first step is to perform a logarithmic colour transform from RGB to HI colour space. Next, a segmentation algorithm extracts the lip area by combining motion with red hue information into a spatio-temporal neighbourhood. The lip's region of interest, semantic information, and relevant boundaries points are then automatically extracted. A good estimate of mouth corners sets active contour initialisation close to the boundaries to extract. Finally, a set of adapted active contours use an open form with curvature

discontinuities along the mouth corners for the outer lip contours, a line-type open active contour when the mouth is closed, and closed active contours with lip shape constrained pressure balloon forces when the mouth is open. They are initialised with the results of the pre-processing stage. An accurate lip shape with inner and outer borders is then obtained with reliable quality results for various speakers under different acquisition conditions.

Chapter VII

This chapter describes a new user-specific 2D to 3D lip animation technique. 2D lip contour position and corresponding motion information are provided from a 2D lip contour extraction algorithm. Static face measurements are obtained from 3D scanners or stereovision systems. The data is combined to generate an initial subject-dependent 3D lip surface. The 3D lips are then modelled as a set of particles whose dynamic behaviour is governed by Smooth Particles Hydrodynamics. A set of forces derived from ellipsoid muscle encircling the lips simulates the muscles controlling the lips motion. The 3D lip model is comprised of more than 300 surface voxels and more than 1300 internal particles. The advantage of the particle system is the possibility of creating a more complex system than previously introduced surface models.

Chapter VIII

The purpose of this chapter is not to describe any lip analysis algorithms but rather to discuss some of the issues involved in evaluating and calibrating labeled lip features from human operators. In the chapter we question the common practice in the field: using manual lip labels directly as the ground truth for the evaluation of lip analysis algorithms. Our empirical results using an Expectation-Maximization procedure show that subjective noise in manual labelers can be quite significant in terms of quantifying both human and algorithm extraction performance. To train and evaluate a lip analysis system one can measure the performance of human operators and infer the "ground truth" from the manual labelers, simultaneously.

Section III
Visual Speech Recognition

This chapter addresses both low and high level problems in visual speech processing and recognition. In particular, mouth region segmentation and lip contour extraction are addressed first. Next, visual speech recognition with parallel support vector machines and temporal Viterbi lattices is demonstrated on a small vocabulary task.

It is well known that visual speech information extracted from video of the speaker's mouth region can improve performance of automatic speech recognizers, especially their robustness to acoustic degradation. However, the vast majority of research in this area has focused on the use of frontal videos of the speaker's face, a clearly restrictive assumption that limits the applicability of audio-visual automatic speech recognition (AVASR) technology in realistic human-computer interaction. In this chapter, we advance beyond the single-camera, frontal-view AVASR paradigm, investigating various important aspects of the visual speech recognition problem across multiple camera views of the speaker, expanding on our recent work. We base our study on an audio-visual database that contains synchronous frontal and profile views of multiple speakers, uttering connected digit strings. We first develop an appearance-based visual front-end that extracts features for frontal and profile videos in a similar fashion. Subsequently, we focus on three key areas concerning speech recognition based on the extracted features: (a) Comparing frontal and profile visual speech recognition performance to quantify any degradation across views; (b) Fusing the available synchronous camera views for improved recognition in scenarios where multiple views can be used; and (c) Recognizing visual speech using a single pose-invariant statistical model, regardless of camera view. In particular, for the latter, a feature normalization approach between poses is investigated. Experiments on the available database are reported in all above areas. To our knowledge, the chapter constitutes the first comprehensive study on the subject of visual speech recognition across multiple views.

Visual speech recognition is able to supplement the information of speech sound to improve the accuracy of speech recognition. A viseme, which describes the facial and oral movements that occur alongside the voicing of a particular phoneme, is a supposed basic unit of speech in the visual domain. As in phonemes, there are variations for the same viseme expressed by different persons or even by the same person. A classifier must be robust to this kind of variation. In this chapter, we describe the Adaptively Boosted (AdaBoost) Hidden Markov Model (HMM) technique (Foo, 2004; Foo, 2003; Dong, 2002). By applying the AdaBoost technique to HMM modeling, a multi-HMM classifier that improves the robustness of HMM is obtained. The method is applied to identify context-independent and context-dependent visual speech units. Experimental results indicate that higher recognition accuracy can be attained using the AdaBoost HMM than that using conventional HMM.

Chapter XII

The basic building blocks of visual speech are the visemes. Unlike phonemes, the visemes are, however, confusable and easily distorted by the contexts in which they appear. Classifiers capable of distinguishing the minute difference among the different categories are desirable. In this chapter, we describe two Hidden Markov Model based techniques using the discriminative approach to increase the accuracy of visual speech recognition. The approaches investigated include Maximum Separable Distance (MSD) training strategy (Dong, 2005) and Two-channel training approach (Dong, 2005; Foo, 2003; Foo, 2002) The MSD training strategy and the Two-channel training approach adopt a proposed criterion function called separable distance to improve the discriminative power of an HMM. The methods are applied to identify confusable visemes. Experimental results indicate that higher recognition accuracy can be attained using these approaches than that using conventional HMM.

Chapter XIII

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.

The aim of this chapter is to examine the possibility of extracting prosodic information from lip features. We used two lip feature measurement techniques in order to evaluate the "lip pattern" of prosodic focus in French. Two corpora with Subject-Verb-Object (SVO) sentences were designed. Four focus conditions (S, V, O or neutral) were elicited in a natural dialogue situation. In the first set of experiments, we recorded two speakers of French with front and profile video cameras. The speakers wore blue lipstick and facial markers. In the second set we recorded five speakers with a 3D optical tracker. An analysis of the lip features showed that visible articulatory lip correlates of focus exist for all speakers. Two types of patterns were observed: absolute and differential. A potential outcome of this study is to provide criteria for automatic visual detection of prosodic focus from lip data.

The information in optical speech signals is phonetically impoverished compared to the information in acoustic speech signals that are presented under good listening conditions. But high lipreading scores among prelingually deaf adults inform us that optical speech signals are in fact rich in phonetic information. Hearing lipreaders are not as accurate as deaf lipreaders, but they too demonstrate perception of detailed optical phonetic information. This chapter briefly sketches the historical context of and impediments to knowledge about optical phonetics and visual speech perception (lipreading). We review findings on deaf and hearing lipreaders. Then we review recent results on relationships between optical speech signals and visual speech perception. We extend the discussion of these relationships to the development of visual speech synthesis. We advocate for a close relationship between visual speech perception research and development of synthetic visible speech.

Section IV
Visual Speaker Recognition

We present a multimodal speaker identification system that integrates audio, lip texture and lip motion modalities, and we propose to use the "explicit" lip motion information that best represent the modality for the given problem. Our work is presented in two stages: First, we consider several lip motion feature candidates such as dense motion features on the lip region, motion features on the outer lip contour, and lip shape features. Meanwhile, we introduce our main contribution, which is a novel two-stage, spatial-temporal discrimination analysis framework designed to obtain the best lip motion features. In speech recognition, the best lip motion features provide the highest phoneme/word/phrase recognition rate, whereas for speaker identification, they result in the highest discrimination among speakers. Next, we investigate the benefits of the inclusion of the best lip motion features for multimodal recognition. Audio, lip texture, and lip motion modalities are fused by the reliability weighted summation (RWS) decision rule, and hidden Markov model (HMM)-based modeling is performed for both unimodal and multimodal recognition. Experimental results indicate that discriminative grid-based lip motion features are proved to be more valuable and provide additional performance gains in speaker identification.

The present chapter reports on the use of lip motion as a stand alone biometric modality as well as a modality integrated with audio speech for identity recognition using digit recognition as a support. First, we estimate motion vectors from images of lip movements. The motion is modeled as the distribution of apparent line velocities in the movement of brightness patterns in an image. Then, we construct compact lip-motion features from the regional statistics of the local velocities. These can be used as alone or merged with audio features to recognize identity or the uttered digit. We present person recognition results using the XM2VTS database representing the video and audio data of 295 people. Furthermore, we present results on digit recognition when it is used in a text prompted mode to verify the liveness of the user. Such user challenges have the intention to reduce replay attack risks of the audio system.