Chapter X
Visual Speech Recognition Across Multiple Views

Patrick Lucey
Queensland University of Technology, Australia

Gerasimos Potamianos
IBM T. J. Watson Research Center, USA

Sridha Sridharan
Queensland University of Technology, Australia

ABSTRACT

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, the authors 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 their recent work. The authors base their study on an audio-visual database that contains synchronous frontal and profile views of multiple speakers, uttering connected digit strings. They first develop an appearance-based visual front-end that extracts features for frontal and profile videos in a similar fashion. Subsequently, the authors 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. This chapter constitutes the first comprehensive study on the subject of visual speech recognition across multiple views.
INTRODUCTION

Recent algorithmic advances in the field of automatic speech recognition (ASR) together with progress in technologies such as speech synthesis, natural language understanding, and dialog modeling have allowed deployment of many automatic systems for human-computer interaction. Of course, these systems require highly accurate ASR to achieve successful task completion and user satisfaction. Although this in general is attainable in relatively quiet environments and for low- to medium-complexity recognition tasks, ASR performance degrades significantly in noisy acoustic environments, especially under conditions mismatched to training data (Junqua, 2000).

One possible avenue proposed for improving ASR robustness to noise is to incorporate visual speech information extracted from a speaker’s face into the speech recognition process – thus giving rise to audio-visual ASR (AVASR) systems. Indeed, over the past two decades, significant progress has been achieved in this field, and many researchers have been able to demonstrate dramatic gains in bimodal ASR accuracy, in line with expectations from human speech perception studies (Sumby and Pollack, 1954). Overviews of such efforts can be found in Chibelushi et al. (2002) and Potamianos et al. (2003), among others. In spite however of this progress, practical deployments of AVASR systems have yet to emerge. This we believe is mainly due to the fact that most research in this field has neglected addressing robustness of the AVASR visual front-end component to realistic video data. One of the most critical overseen issues is speaker head pose variation, or in other words the camera view-point of the speaker’s face.

Indeed, with a few exceptions reviewed in the Background section, nearly all work in the literature has concentrated on the case where the speaker’s face is captured in a fully frontal pose – a rather restrictive human-computer interaction scenario, a fact also made clear in Figure 1. For example, one potential AVASR application is speech recognition using mobile devices such as cell phones. Device placement with respect to the head does not allow frontal AVASR in this case. Another interesting scenario is that of in-vehicle AVASR. Due to frequent driver head movement, a frontal pose cannot be guaranteed, regardless of camera placement – for example at the rear-view mirror, the cabin driver-side column, or the instrument console. Other possibilities include the design of an audio-visual headset, where a miniature camera is placed next to the microphone in the wearable boom. Requiring frontal views of the speaker mouth means that the device may be designed to protrude unnecessarily in front of the mouth, creating headset instability and usability issues (Gagne et al., 2001; Huang et al., 2004). In contrast, placing the camera to the side of the face would allow a significantly shorter boom, hence resulting in a lighter and easier to use headset. Finally, an interesting scenario is this of AVASR during meetings and lectures inside smart rooms. There, pan-tilt-zoom (PTZ) cameras can track the meeting speaker(s) providing high resolution views. However, due to the camera fixed placements in space, frontal speaker views cannot be guaranteed. This latter scenario motivates our work. It is discussed in more detail later on in this book chapter, together with the audio-visual database collected in this domain to drive our research.

Motivated by the above, in this chapter, we advance beyond the single-camera, frontal-view AVASR paradigm that has hindered progress in the field. Since the subject of non-frontal AVASR is still at its infancy, we focus on a simplified version of the problem, by considering two fixed views: the traditional frontal view and an extreme case, namely that of profile views. To allow a comprehensive study of the relevant issues, we consider the case where data are available synchronously from both views via a multi-camera setting (see also Figure 2). In addition, in this study we bypass the problem of head-pose
Related Content

Effects of a Technology-Enriched, Task-Based Language Teaching Curriculum on Chinese Elementary Students’ Achievement in English as a Foreign Language
[www.igi-global.com/chapter/effects-of-a-technology-enriched-task-based-language-teaching-curriculum-on-chinese-elementary-students-achievement-in-english-as-a-foreign-language/108782?camid=4v1a](www.igi-global.com/chapter/effects-of-a-technology-enriched-task-based-language-teaching-curriculum-on-chinese-elementary-students-achievement-in-english-as-a-foreign-language/108782?camid=4v1a)

Demystifying Domain Specific Languages
[www.igi-global.com/chapter/demystifying-domain-specific-languages/108723?camid=4v1a](www.igi-global.com/chapter/demystifying-domain-specific-languages/108723?camid=4v1a)

A Generic Approach for the Semantic Annotation of Conceptual Models Using a Service-Oriented Architecture
[www.igi-global.com/chapter/a-generic-approach-for-the-semantic-annotation-of-conceptual-models-using-a-service-oriented-architecture/108788?camid=4v1a](www.igi-global.com/chapter/a-generic-approach-for-the-semantic-annotation-of-conceptual-models-using-a-service-oriented-architecture/108788?camid=4v1a)

Morphological Analysis of Ill-Formed Arabic Verbs for Second Language Learners
[www.igi-global.com/chapter/morphological-analysis-ill-formed-arabic/61060?camid=4v1a](www.igi-global.com/chapter/morphological-analysis-ill-formed-arabic/61060?camid=4v1a)