Audiovisual Facial Action Unit Recognition using Feature Level Fusion

Zibo Meng, University of South Carolina, Columbia, SC, USA
Shizhong Han, University of South Carolina, Columbia, SC, USA
Min Chen, University of Washington Bothell, Bothell, WA, USA
Yan Tong, University of South Carolina, Columbia, SC, USA

ABSTRACT
Recognizing facial actions is challenging, especially when they are accompanied with speech. Instead of employing information solely from the visual channel, this work aims to exploit information from both visual and audio channels in recognizing speech-related facial action units (AUs). In this work, two feature-level fusion methods are proposed. The first method is based on a kind of human-crafted visual feature. The other method utilizes visual features learned by a deep convolutional neural network (CNN). For both methods, features are independently extracted from visual and audio channels and aligned to handle the difference in time scales and the time shift between the two signals. These temporally aligned features are integrated via feature-level fusion for AU recognition. Experimental results on a new audiovisual AU-coded dataset have demonstrated that both fusion methods outperform their visual counterparts in recognizing speech-related AUs. The improvement is more impressive with occlusions on the facial images, which would not affect the audio channel.

KEYWORDS
Action Units, Convolutional Neural Network, Facial Action Unit Recognition, Facial Activity, Feature-Level Information Fusion

INTRODUCTION
Facial activity is one of the most powerful and natural means for human communication (Pantic & Bartlett, 2007a). Driven by the recent advances in human-centered computing, there is an increasing need for accurate and reliable characterization of the displayed facial behavior. The Facial Action Coding System (FACS) developed by Ekman and Friesen (Ekman, Friesen, & Hager, 2002) is the most widely used and objective system for facial behavior analysis. Based on the FACS, the facial behavior is described by a small set of facial Action Units (AUs), each of which is anatomically related to the...
contraction of a set of facial muscles. Given different interpretation rules or systems, e.g. Emotion FACS rules (Ekman et al., 2002), AUs have been used in inferring various human affective states. In addition to the application of human behavior analysis, an automatic system for facial AU recognition is desired in interactive games, online/remote learning, and other human computer interaction (HCI) related applications.

As demonstrated in the survey papers (Pantic, Pentland, Nijholt, & Huang, 2007b; Zeng, Pantic, Roisman, & Huang, 2009; Sariyanidi, Gunes, & Cavallaro, 2015), great progress has been made over the years on automatic AU recognition from posed/deliberated facial displays. Recognizing facial AUs from spontaneous facial displays, however, is challenging due to subtle and complex facial deformation, frequent head movements, temporal dynamics of facial action, etc. Furthermore, it is especially challenging to recognize AUs involved in speech. As discussed in (Ekman et al., 2002), the AUs are usually activated at low intensities with subtle facial appearance/geometrical changes when they are responsible for producing speech. In addition, they will often introduce ambiguity, e.g., occlusions, in recognizing other AUs.

For example, pronouncing a phoneme /b/ has two consecutive phases, i.e., Stop and Aspiration phases. In the Aspiration phase, the lips are apart and the oral cavity between the teeth is visible, as shown in Figure 1(b), which are the major facial appearance clues to recognize AU25 (lips part) and AU26 (jaw drop), respectively. In the Stop phase, the lips are pressed together due to the activation of AU24 (lip presser), as shown in Figure 1(a). Consequently, the oral cavity is occluded by the lips and AU26 is “invisible” in the visual channel.

All existing approaches on facial AU recognition extract information solely from the visual channel. In contrast, this paper proposes a novel approach, which exploits the information from both visual and audio channels, to recognize speech-related AUs. This work is motivated by the fact that facial AUs and voice are highly correlated in natural human communications. Specifically, voice/speech has strong physiological relationships with some lower face AUs such as AU25 (lips part), AU26 (jaw drop), and AU24 (lip presser) because jaw and lower-face muscle movements together with the soft palate, tongue and vocal cords produce the voice.

These relationships are well recognized and have been exploited in natural human communications. For example, without looking at the face, people will know that the other person is opening his/her mouth when hearing laughter. Following the example of recognizing AU26 (jaw drop) in the Stop phase of pronouncing the phoneme /b/, we can infer that AU26 (jaw drop) has been activated when hearing the sound /b/, even when it is “invisible” in the visual channel.

Figure 1. Example images of speech-related facial behaviors, where different combinations of AUs are activated to pronounce a phoneme /b/
Semi-Supervised Multimodal Fusion Model for Social Event Detection on Web Image Collections
Zhenguo Yang, Qing Li, Zheng Lu, Yun Ma, Zhiguo Gong, Haiwei Pan and Yangbin Chen (2015). *International Journal of Multimedia Data Engineering and Management* (pp. 1-22).
www.igi-global.com/article/semi-supervised-multimodal-fusion-model-for-social-event-detection-on-web-image-collections/135514?camid=4v1a

Customizing Digital Storefronts Using the Knowledge-Based Approach
www.igi-global.com/chapter/customizing-digital-storefronts-using-knowledge/22952?camid=4v1a

The Haunted School on Horror Hill: A Case Study of Interactive Fiction in an Elementary Classroom
www.igi-global.com/chapter/haunted-school-horror-hill/49457?camid=4v1a