Chapter IV

Visual Speech and Gesture Coding Using the MPEG-4 Face and Body Animation Standard

Eric Petajan
VectorMAX Corporation, USA

ABSTRACT

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.

INTRODUCTION

In recent years the number of people accessing the internet or using digital devices has exploded. In parallel the mobile revolution is allowing consumers to access the internet on relatively powerful hand-held devices. While the transmission and display of information is efficiently handled by maturing fixed and wireless data networks and terminal devices, the input of information from the user to the target system is often impeded by the lack of a keyboard, low typing skills, or busy hands and eyes. The last barrier to efficient man-machine communication is the lack of accurate speech recognition in real-world environments. Given the importance of mobile communication and computing, and the ubiquitous
internetworking of all terminal devices, the optimal system architecture calls for compute-intensive processes to be performed across the network. Support for thin mobile clients with limited memory, clock speed, battery life, and connection speeds requires that visual speech and gesture information captured from the user be transformed into a representation that is both compact and computable on the terminal device.

The flow of audio/video data across a network is subject to a variety of bottlenecks that require lossy compression; introducing artifacts and distortion that degrade the accuracy of scene analysis. Video with sufficient quality for facial capture must be either stored locally or analyzed in real time. Real-time video processing should be implemented close to the camera to avoid transmission costs and delays, and to more easily protect the user’s visual privacy. The recognition of the human face and body in a video stream results in a set of descriptors that ideally occur at the video frame rate. The human behavior descriptors should contain all information needed for the Human-Computer Interaction (HCI) system to understand the user’s presence, pose, facial expression, gestures, and visual speech. This data is highly compressible and can be used in a communication system when standardized. The MPEG-4 Face and Body Animation (FBA) standard provides a complete set of Face and Body Animation Parameters (FAPs and BAPs) and a codec for super low bit-rate communication. This chapter describes the key features of the MPEG-4 FBA specification, its application to visual speech and gesture recognition, and architectural implications.

The control of a computer by a human incorporating the visual mode is best implemented by the processing of video into features and descriptors that are accurate and compact. These descriptors should only be as abstract as required by network, storage capacity, and processing limitations. The MPEG-4 FBA standard provides a level of description of human facial movements and skeleton joint angles that is both highly detailed and compressible to 2 kilobits per second for the face and 5-10 kilobits per second for the body. The MPEG-4 FBA stream can be transmitted over any network and can be used for visual speech recognition, identity verification, emotion recognition, gesture recognition, and visual communication with the option of an alternate appearance. The conversion of video into an MPEG-4 FBA stream is a computationally intensive process which may require dedicated hardware and HD video to fully accomplish. The performance of recognition tasks on the FBA stream can be performed anywhere on the network without risking the violation of the users visual privacy when video is transmitted. When coupled with voice recognition, FBA recognition should provide the robustness needed for effective HCI. As shown in Figure 1, the very low bit-rate FBA stream enables the separation of the HCI from higher level recognition systems, applications and databases that tend to consume more processing and storage than is available in a personal device. This client-server architecture supports all application domains including human-human communication, human-machine interaction, and local HCI (non-networked). While the Humanoid Player Client exists today on mobile phones, a mobile Face and Gesture Capture Client is still a few years away.

MPEG-4 FBA streaming is commercially deployed today for delivery of animation to mobile phones in Europe within the AniTonesTM application. Content is efficiently created using facial capture from ordinary video and combined with animated backgrounds and bodies. FAP and BAP data (including both body and background object animation) are compressed into one small package for delivery over any mobile IP network. The AniTonesTM player maintains a frame-rate of fifteen frames per second with synchronized audio on a 109 MHz processor.