IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB7305

Chapter IV

MPEG-4 Facial Animation and its Application to a Videophone System for the Deaf

Nikolaos Sarris and Michael G. Strintzis
Aristotle University of Thessaloniki, Greece

INTRODUCTION

This chapter aims to introduce the potential contribution of the emerging MPEG-4 audio-visual representation standard to future multimedia systems. This is attempted by the 'case study' of a particular example of such a system--'LipTelephone'--which is a special videoconferencing system being developed in the framework of MPEG-4 (Sarris et al., 2000b). The objective of 'LipTelephone' is to serve as a videophone that will enable lip readers to communicate over a standard telephone connection. This will be accomplished by combining model-based with traditional video coding techniques in order to exploit the information redundancy in a scene of known content, while achieving high fidelity representation in the specific area of interest, which is the speaker's mouth. Through this description, it is shown that the standard provides a wide framework for the incorporation of methods that had been the object of pure research even in recent years. Various such methods are referenced from the literature, and one is proposed and described in detail for every part of the system being studied. The main objective of the chapter is to introduce students to these methods for the processing of multimedia material, provide to researchers a reference to the state-of-the-art in this area and urge engineers to use the present research methodologies in future consumer applications.

This chapter appears in the book, Multimedia Networking: Technology, Management and Applications by Syed Mahbubur Rahman.

CONVENTIONAL MULTIMEDIA CODING SCHEMES AND STANDARDS

The basic characteristic and drawback of digital video transmission is the vast amount of data that need to be stored and transmitted over communication lines. For example, a typical black-and-white videophone image sequence with 10 image frames per second and dimensions 176 x 144 needs 2Mbits/sec transmission rate (8 bits/pixel x [176x144] pixels/frame x 10 frames/sec). This rate is extremely high even for state-of-the-art communication carriers and demands high compression, which usually results in image degradation.

In recent years many standards have emerged for the compression of moving images. In 1992 the Moving Picture Experts Group (MPEG) completed the ISO/IEC MPEG-1 video-coding standard, while in 1994 the MPEG-2 standard was also approved (MPEG, online). These standards contributed immensely to the multimedia technological developments as they both received Grammy awards and made interactive video on CD-ROM and digital television possible (Chariglione, 1998). The ITU-T (formerly CCITT) organization established the H.261 standard in 1990 (ITU-T, 1990) and H.263 in 1995 (ITU-T, 1996), which were especially targeted to videophone communications. These have achieved successful videophone image sequence transmission at rates of approximately 64 Kbps (Kbits per second).

The techniques employed in these standards were mainly based on segmentation of the image in uniformly sized 8x8 rectangular blocks. The contents of the blocks were coded using the Discrete Cosine Transform (DCT), and their motion in consecutive frames was estimated so that only the differences in their positions had to be transmitted. In this way spatial and temporal correlation is exploited in local areas and great compression rates are achieved. The side effects of this approximation, however, are visual errors on the borders of the blocks (blocking effect) and regularly spaced dots on the reconstructed image (mosquito effect). These effects are more perceptible when higher compression rates are needed, as in videophone applications. In addition, these standards impose the use of the same technique on the whole image, making it impossible to distinguish the background or other areas of limited interest, even when they are completely still (Schafer & Sikora, 1995; IMSPTC, 1998).

These problems are easily tolerated during a usual videoconference session where sound remains the basic means of communication, but the system is rendered useless to potential hearing-impaired users who exploit lip reading for the understanding of speech. Even for these users however, image degradation could be tolerated in some areas of the image. For example, the background does not need to be continuously refreshed, and most areas of the head, apart from the mouth area, do not need to be coded with extreme accuracy. It is therefore obvious that multimedia technology at that time could benefit immensely from methods that utilize knowledge of the scene contents, as these could detect different objects in the scene and prioritize their coding appropriately (Benois et al., 1997).

THE MPEG-4 STANDARD

In acknowledgment of the previously mentioned limitations of the existing standards, MPEG launched in 1993 a new standard called MPEG-4 which was approved in Version 1 in October 1998 and in Version 2 in December 1999. MPEG-4 is the first audio-visual representation standard to model a scene as a composition of objects with specific