EXPECTATION OF FUTURE SERVICES IN MOBILE MULTIMEDIA COMMUNICATION

T. S. Huang
Beckman Institute for Advanced Science and Technology
University of Illinois
405 N. Mathews Avenue, Urbana, IL 61801, U. S. A.
Ph. 217-244-1638, FAX 217-244-1642
E-mail huang@uicsl.csl.uiuc.edu

K. Aizawa
Department of Electrical Engineering,
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
Ph. 03-3812-2111, FAX 03-3818-5706
E-mail aizawa@hal.t.u-tokyo.ac.jp

I. What is Multimedia?

Everyone has his own definition of the term "multimedia". Ours is as follows. First of all, we consider communication systems in a very broad sense. They include commnication between man and machine as well as communication between men via machine. An example of the former is computer-aided learning; an example of the latter is teleconferencing. A system is a multimedia system, if it satisfies three criteria:

(a) It involves several human senses (at least vision and hearing) and several signal modalities (graphics, still images, video; speech, music, other sounds; ...).

(b) It is human-machine interactive.

(c) The machine has a hyper data structure, so that the retrieval of information is very flexible.

II. Mobile Communication

We shall concentrate on the case of mobile video phones. Normally it will be used to transmit scenes of human faces. However, in many cases we

will also like to use them to transmit documents/graphics, still images, and video of scenes other than faces (perhaps at a reduced frame rate). With the low available bandwidth, it is unrealistic to expect to be able to transmit images at high resolution and high frame rate. However, we do hope that the system would provide the flexibility of trading off between image resolution and frame rate.

We now restrict our attention to the transmission of scenes of human faces. We would like to squeeze the video of a face (talking and expressing emotions) through a channel of 5-10 Kits/second. At this bit rate, we can hardly expect to get high quality video. But, on the other hand, the expectation is also low so that perhaps something could be worked out. This is an application where model- based compression methods may play a key role.

III. Computer Vision and Very Low Bitrate Video Compression

How could we achieve a bit rate of 5-10 Kbits/second or lower for video of human faces? We believe ideas and techniques from Computer vision could contribute a great deal. Let us look at some possibilities.

(a) Segmentation and feature extraction:

Researchers in computer vision have been working on the two related problems of segmentation and feature extraction for a long time, in some cases, specifically aiming at segmenting out faces and extracting facial features such as eyes, and mouth. Techniques in edge detection, active contours, elastic templates, etc., as well as the use of color and motion can be very useful. There are many potential applications of segmentation and feature extraction to video compression. We can segment out faces and code them defferently from the background. Facial feature extraction is essential in some of the main approaches to model-based coding.

(b) 3D shape from 2D shading, texture, etc.:

To be able to get 3D geometrical information of the face and lighting conditions will help enormously the realistic synthesis of video sequences in model-based coding. Experiments have shown that just updating the geometry of

the face does not give good synthesized face sequences. On the other hand, updating texture is often too expensive in terms of the additional bits needed. Lighting conditions and 3D face geometry can be used to help texture updating.

(c) Motion estimation:

Motion compensation is a key component of most video compression and interpolation methods. However, almost all current motion compensation methods estimates only 2D displacements. Computer vision can help in a number of ways. First, many of the optic flow estimation techniques may be useful in better determining the 2D displacement vectors. Second, Techniques for estimating 3D rigid and nonrigid motion can potentially improve the compression factor and the reconstructed image quality considerably. Third, methods of handling multiple motion including transparency can help motion estimation both in 2D and in 3D.

(d) Recognition:

Perhaps the best hope in achieving really low bitrate (1 Kbits per second or lower) is to transmit the facial movement information at the semantic level. If we allow for 16 emotional expressions (happy, sad, surprized, etc.) and 128 phonemes, and assume we update every 1/10 th of a second, then for facial movement information, we need only 110 bits/second. Of course, based on this information, we can synthesized only rather crude facial sequences. But in the context of mobile video phone, the quality may be acceptable. Futuremore, recognizing expressions and phonemes is probably much easier than estimating the numerical values of the displacement vectors of the key feature points on the face. In fact, the transmitted speech can be recognized and used to drive the face model at the receiving end, eliminating the need to do computer lip reading.

IV. Concluding Remarks

We advocate that a mobile video communication system should be flexible enough to permit the transmission of not only conversational video but also high-resolution documents/graphics and still images (of course at a much

slower rete). Thus, the ability to trade off between image resolution and frame rate is highly desirable.

For video of face scenes, the use of computer vision techniques (especially recognition) appears to be a good approach toward achieving very low bitrate communication (5-10 Kbits/second or even lower).

Mobile video communication is a very important application of MPEG4. Thus researchers in this field should be actively involved in shaping the MPEG4 decisions. In particular, the requirements of mobile video communication should be clarified and crystalized, and conveyed to the MPEG4 group. Techniques for fulfilling these requirements should be studied and incorporated in the final MPEG4 recommendations.

Acknowledgement

In formulating ideas expressed in this paper, we were benefitted tremendously from discussions with Professor Rama Chellappa, Department of Electrical Engineering, University of Maryland.