Focus of Attention: Towards Low Bitrate Video Tele-Conferencing

Jie Yang Leejay Wu Alex Waibel

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

Low bitrate video tele-conferencing requires adapting algorithms that may work perfectly well in a high-bitrate situation. When a slow transmission rate is unacceptable, compromise must be reached among the demands of speed, bandwidth limits and image quality. In this paper we present an approach to low bitrate video teleconferencing by focusing attention on important information. We show that by selectively degrading the quality of less important regions, more important regions can be sent without loss of quality but with greatly reduced bandwidth requirements. A prototype system has been developed to demonstrate the concept. The experimental results show significant savings of required bandwidth for video subjected to the changes.

1 Introduction

The demand for video communication between geographically distant sites has increased greatly in the last decade. Applications depending on video delivery include video teleconferencing and telephony, interactive multimedia and multimedia e-mail. Although rapid progress has been made in technologies of digital communications, demand for data transmission bandwidth is too much for current systems. Very low bitrate tele-conferencing is another research field for addressing such a problem [1].

Video delivery depends not only on the bandwidth and data size, but also the network traffic. While sending the same amount of data via even the same network, different network traffic can result in different delivery times. For example, video transmission over a wireless network requires not only adaptation to changes in bandwidth and channel characteristics, but also the amount of other traffic. In a video conference, it might be more important to keep updating the most interesting information than slowing to send the less interesting data. Partial or lower quality real-time images might be preferable to complete but delayed images. It is desirable for a video conference system to optimally keep sending the most important information at a reasonable speed based on network characteristics.

In this paper we present an approach to low bitrate video teleconferencing by focusing attention on important information. The key idea is to use an object-centered representation to extract the most important information in video image and then send only that through the network. In this way a one can make better use of available network resources to achieve optimal performance without losing important information. The implementation is to add a selection function on top of classical coding system and make no change on the original coder and network protocols. This new approach differs from the model-based coding approach [2] in that no model is needed at the receiver.

In order to demonstrate the proposed approach, we have developed a face tracker-based tele-conferencing system. The system can automatically focus its attention on a given face found by a real-time face-tracker [3] and its adjustable surrounding area, which is then fed to the coding system and sent to the receiver. The face can be selected by a mouse click or finger pointing if a touch screen is used. Based on the information provided by the face tracker and the network traffic, a window surrounding the face can be determined. The window size can reflect the network traffic. The image outside of the window will be either cropped or blurred. The preprocessed image is then fed to a tele-conferencing software package- *vic*, a real-time multimedia application for video conferencing over the Internet [4]. Some modifications have been made to make it send the selected images. The system has been successfully running in our lab across different Alpha and HP machines. We have performed several experiments to evaluate the system performance. The experiments have shown some significant measurement results.

2 Focus of Attention

Rapid progress in digital communications systems performance and mass-storage density has provided opportunities for new network-based multimedia applications such as video teleconferencing, video telephony, interactive multimedia and multimedia e-mail. Performance over current packet-switched networks such as the Internet, however, can degrade significantly when the network is overloaded. Coding is another way to minimize the bandwidth required for transmission of high quality/rate video data, as well as the storage capacity required for saving such data in fast storage media and in archival databases. Video coding (or compression) is the process of reducing the number of bits required to represent video data subject to a fidelity criterion on the final representation.

Video delivery depends on the codec and network bandwidth as well as network traffic. Reliance on increasing bandwidth and data compression is not sufficient, as both may be exceed current capabilities or be unreasonably costly. The problem becomes more complex when the network cannot provide network services with performance guarantees. Therefore, one of the most challenging problems in multimedia communications is how to keep video streams at a reasonable level of quality when the network cannot provide performance guaranteed service. We present a solution to such a problem by adding a selection module on the top of a codec as shown in Figure 1. The technique is the use of feature-indicating interest images to focus attention on specific areas of the video imagery. The motivation comes from scientific studies on selective visual attention [5].

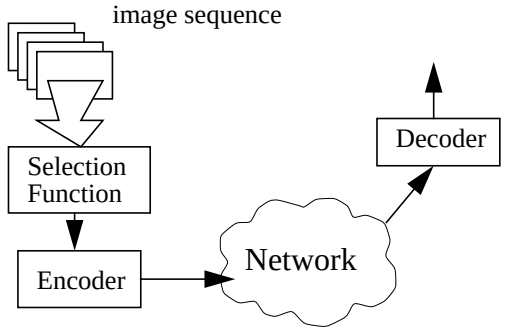

Figure 1 Focus attention for video transmission

Humans outperform computers in many pattern recognition and navigation tasks. One explanation is that computers have to process all available information, whereas a human quickly focuses attention to the most important information without paying too much to less relevant information. The limited processing capacity of humans has imposed the need for selectivity in processing [6] [7]. The same strategy can be used by a machine when it is subject to the same limitations as a human. Focus of attention has been investigated in a variety of contexts. One of the largest branches of study has studied attention in static images [8] [9]. The concept of "spotlight" was used to explain how information is excluded from further analysis. Computational models of the spotlight mechanism have been implemented using neural networks [10] [11]. For a sequence of images, the process of focusing attention becomes more challenging because objects can move and change.

Many studies on focus of attention have directed to selecting relevant inputs to improve performance and robustness of a system. In this study, the concept of focus of attention is used to select important information for achieving optimal performance under the constraints of limited bandwidth and varying network traffic. In a video tele-conference, not all information is equally important. Using an object-centered representation, it is possible to focus its attention on and send only the most important information. People are important objects in a tele-conference; thus, we can partition the scene into several sub-scenes based on each individual's position and then select sub-scenes to send based on their priorities and network traffic. For example, if we are interested in a speaker, the speaker's face and surrounding area can be contain the most important information, and have the highest priority to be sent to the receiver.

The mechanism of selection and data discarding is similar to that of the human eye, which is not equipped with a uniform resolution over the whole visual field. Near the optical axis it has the fovea where the resolution (over a one degree range) is higher by an order of magnitude than that in the periphery. A human can view a large visual field by moving the fovea rapidly.

3 Tracking Human faces in Real-Time

A human face provides a variety of different communicative functions such as identification, perception of emotional expressions, and lip-reading. To track a human face, the system needs to be able to not only locate a face, but also find the same face throughout a sequence of images. This requires the system to have the ability to estimate the motion while locating the face. Furthermore, to track faces outside a certain range the system needs to control the camera, e.g., panning, tilting, and zooming. We have developed a real-time face tracker [4]. The camera's panning, tilting, and zooming are controlled by the program. Analog video images are digitized by a framegrabber into RGB values. Three types of models have been employed in developing the system. First, a stochastic model is used to characterize skin-colors of human faces. The information provided by the model is sufficient for tracking a human face in various poses and views. This model can adapt in real-time to different people and different lighting conditions. Second, a motion model is used to estimate image motion and therefore determine a search window. Third, a camera model is used to predict and to compensate for camera motion. The face tracker has achieved a rate of 30+ frames/second using a workstation with a framegrabber and a video camera. It can track a person's face while the person moves freely (e.g., walking, jumping, sitting and rising) in a room as shown in Figure 2.

Figure 2 An example of tracking a face

4 A Prototype System

In order to demonstrate the proposed approach, we developed a prototype system. The chosen video-conferencing application was Vic-2.6, by Steve McCanne (mccanne@ee.lbl.gov) and Van Jacobson (van@ee.lbl.gov) as one of a suite of multimedia conferencing utilities. Vic provides the video interface by accepting video data and transmitting it to recipients who in turn may be sending back video. In this setup, a DEC Alpha equipped with a J300 card and a VC-C1 Canon camera was directly supported with several session encoding methods -- nv, CellB, scr, JPEG, and H261, each of which could be used to correspond with another program that supported the appropriate format. The video data for the nv, CellB, and scr sessions all store the video data in YUV format in the same type of buffer; hence, modifications to the data at that stage affect these three types.

The source code for the jv300 grabber routines was modified to use sockets-based communication with the server to check for new coordinates of a region to target. The rectangle thus selected could be subjected to a number of modifications, depending on the options selected by the user via a Tcl/Tk script. All modifications were fairly gross and simple, for the sake of speed.

The target area is chosen by the selection function based on the face tracker. Each modification takes place right before the target frame is encoded. A function is used to obtain coordinate information, read the configuration file, and call the appropriate image editing routines before passing the data to the encoder.

Three modes have been implemented in the prototype system as shown in Figure 3 and they can be switched at any time:

Pseudo-cropping: This option causes the entire area outside the selected rectangle to be turned black, saving a significant amount of transfer time, as fewer pixels change between frames and thus the delta-based encoding of each session type can send fewer pixels. At the cost of aesthetics and a large portion of the image, this provides fairly significant savings with little cost in processing.

Slicing: This option causes the 25% of the image including and around the facial region to be sent as a smaller frame image, with the rest being discarded. Unlike the previous option, this choice always discards the same amount of data, and there is minimal loss in appearance, as no editing is evident on the receiving end. However, significant motion of the head will cause it to appear as if the camera were panning or tilting, and this saves slightly less bandwidth at a slightly greater cost in processing time.

Blurring: This option causes the area outside the facial region to undergo an averaging process, with coarseness set by the user. The background is divided into a number of rectangles, each of which is filled with the average values of the component pixels. Even at the very low levels of coarseness, where the loss of detail is noticeable but not particularly bad, the savings in bandwidth tend to be considerable. At the highest levels, the background becomes almost completely uninformative and the savings approximate those of simply discarding the background as in the pseudo-cropping method.

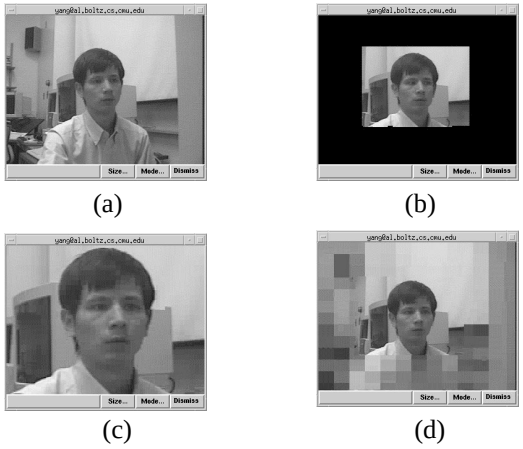

Figure 3 An example of an image in different modes: (a) original; (b) pseudo-cropping; (c) slicing; (d) blurring

In a video-conferencing session, there can be multiple people on each end, and one party might wish to track the current speaker in the other, regardless of that speaker's relative position to the camera. By default, the face-tracking algorithm selects the largest face to start tracking. If two faces are present, and one is significantly closer to the camera than the other, the closer face will be chosen (assuming that the faces are of equal physical size). The tracker itself does not directly provide the controls for choosing faces, but does allow any point to be used as a focal point for the search. We have developed a user interface to allow the user to select a place to start searching.

The interface is a Tcl/Tk-based program used for specifying focal points for the face-tracker's search algorithm. It includes a window to allow the user to select a point; a pair of buttons that controls whether the input device is the mouse or a touch-screen; and buttons for panning, tilting, focusing, and zooming the camera. The implementation is twofold; the GUI routines are written in Tcl/Tk, but the Tcl/Tk script is controlled by C code, which uses the Tcl/Tk code to draw everything and check for mouse input, but then controls the actual effects after translating the mouse clicks and screen touches to coordinates.

5 Experimental Results

To test the proposed method, we performed various experiments. The system has been running on different platforms such as Alpha and HP workstations in our lab. To minimize the effects of network traffic, we compared the total frame sizes transmitted by vic according to its own count. Two sequences of 100 frame images were collected for such comparison. Two versions were in testing; one in which it behaved normally but also recorded the unmodified YUV image data to a file, along with the corresponding facial coordinates, and a second version which replaced the camera and socket input with that read from the YUV and coordinate file. These were used to insure that in two trial sets, every modification used the exact same frames and face coordinates.

Figure 4 shows normalized averages of Trials 1 and 2; it graphs the results and average of the modifications, including different degrees of blurring, versus total size of frames. Several interesting observations have been discovered in experiments. There are fairly significant differences in the results of the two trials. This can generally be explained in terms of the differing face-to-frame-size ratios; the closer the face to the camera and thus the higher ratio, the less background there is to reduce, and therefore less reduction of frame size from background-oriented methods. Movement also matters, since the changes in the background covered or revealed affect the different modifications to different degrees.

The savings in frame sizes through slicing were close to identical, about 50% when sending 25% of the image; here, as the program sent the subimages of constant size, there were fewer possible variations.

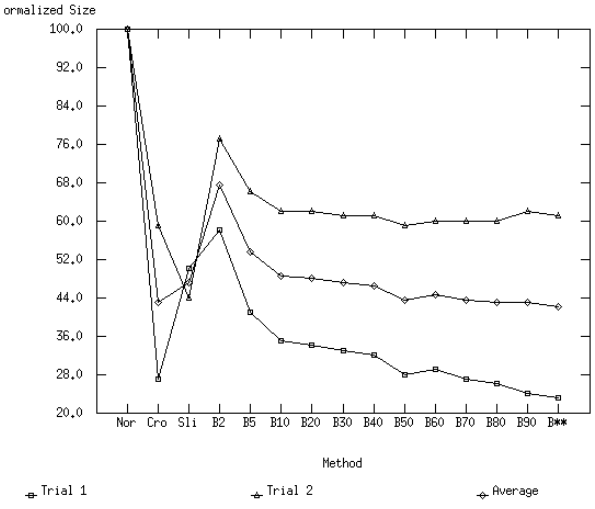

Figure 4 Averages of Trials 1 and 2 (Equal Weights)

In both trials, blurring and cropping provided marked decreases in the frame size, with the higher levels of blurring essentially sacrificing a recognizable background to quickly send whatever region is designated interesting (i.e. that denoted by the coordinates from the face tracker). A blurring intensity gradient might instead be feasible, to keep some of the image quality in the vicinity but retain most of the savings in bytes transferred.

6 Conclusion

We have demonstrated that a selection mechanism on the top of a normal codec to help choose particular regions can yield significant savings in the required bandwidth. Thus, in a video-conferencing situation where low bandwidth is a more restrictive constraint than computing power, the use of such measures as blurring or even outright discarding the data outside an important region can preserve the most significant information to minimize transfer requirements, all without requiring a special client on the receiving side.

7 Acknowledgments

This research was sponsored by the Advanced Research Projects Agency under the Department of the Navy, Naval Research Office under grant number N00014-93-1-0806.

References

[1] H. Li, A. Lundmark, and R. Forchheimer, “Image sequence coding at very low bit rates: a review,” IEEE Transactions on Image Processing, Vol.3, No.5, pp. 589-609, 1994.

[2] K. Aizawa and T.S. Huang, “Model-based image coding advanced video coding techniques for very low bit-rate applications,” Proceedings of the IEEE, Vol. 83, No.2, pp. 259-71, 1995.

[3] J. Yang and A. Waibel, “Tracking human faces in real-time,” CMU CS Technical Report, CMU-CS-95-210, November, 1995

[4] S. McCanne and V. Jacobson, “vic: a flexible framework for packet video,” Proceedings of ACM Multimedia ‘95.

[5] C. Bundesen, “A theory of visual attention,” Psychological Review, Vol. 97, No. 4, pp. 523-547, 1990.

[6] D.E. Broadbent, “Task combination and selective intake of information,” Acta Psychologica, Vol. 50, pp. 253-290, 1982.

[7] A. Allport, “Visual attention,” in: Posner, M., ed., Foundations of Cognitive Science, MIT Press, Cambridge, Ma., pp. 631-683, 1989.

[8] A. Triesman, & G. Gelade, “A feature integration theory of attention,” Cognitive Psychology: Vol. 12, pp. 97-136, 1980.

[9] A. Hulbert and T. Poggio, “Spotlight on attention,” MIT AI Laboratory Memo AI- 817. Massachusetts Institute of Technology, Cambridge, MA, 1985.

[10]C. Koch, and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” Human Neurobiology, Vol. 4, pp. 219-227, 1985.

[11]M.C. Mozer, “A connectionist model of selective attention in visual perception,” Technical Report CRG-TR-99-4, University of Toronto, Toronto, Canada, 1988.