REGION OF INTEREST PRIORITY CODING FOR SIGN LANGUAGE VIDEOCONFERENCING

Richard P. Schumeyer[1] **Edwin A. Heredia**[2] **Kenneth E. Barner**[1]
[1]Applied Science and Engineering Laboratories,
P.O. Box 269, Wilmington, DE 19899 USA
Voice: (302)651-6841 Fax:(302)651-6895
{schumeye,barner}@asel.udel.edu
[2]Thomson Consumer Electronics, Corporate Innovation and Research,
INH-700, P.O. Box 6139 Indianapolis, IN 46206-6139 USA
HerediaE@rnd3.indy.tce.com

Abstract - This paper investigates compression of sign-language video sequences for transmission over low bitrate channels. The motion of hands and arms inherent in sign language requires high temporal resolution and greater compression for band-limited channels. Most existing compression schemes treat the entire image uniformly. However, the background is much less critical than the foreground for sign language communication. A new algorithm is developed that gives priority to the foreground and sacrifices the quality of the background to achieve higher compression rates and higher foreground quality without compromising the frame rate. The algorithm is extended to use a dynamic region of interest defined by face and hand location. Results show that the frame rate and quality of the region of interest are higher than with uniform algorithms.

1 INTRODUCTION

Sign language is the medium of choice for in-person communication between deaf speakers. Sign language is fast: propositions may be communicated in sign language in about the same amount of time as speech[1]. Signing is also expressive. Body posture, facial expression, manner of articulation, etc., perform the same function in signed communication as prosody in speech.

Sign language communication between remote individuals has been generally unavailable, due to a lack of affordable systems capable of transmitting two-way video. Instead, text systems such as TDDs were the only convenient choice for remote communication between hearing-impaired individuals. TDD is inferior to sign language in both rate and expressiveness. TDD speed is limited by typing ability, which results in slower communication than via signing.

The increasing availability of affordable communication channels such as ISDN and the Internet, together with readily available videoconferencing software compatible with established standards, offers this population the possibility of remote sign language communication. These newly available channels have low bit rates. BRI-ISDN operates at 128 kbps and the Internet has variable throughput that depends on congestion and infrastructure. The new channels provide the ability to transmit at

rates sufficient for video transmission, at prices low enough to be within reach of many consumers.

To date, most videoconferencing research has been tested on head and shoulders sequences. Sign language, which features moving arms and hands, contains more motion than a typical talking head sequence. The additional motion, together with the requirement for a frame rate high enough for smooth motion perception, results in increased bandwidth for predictive compression schemes. This paper develops a coding algorithm for sign language which defines a region-of-interest (ROI) and allocates more bandwidth to the ROI. Both static and dynamic ROIs are explored.

Video transmission standards such as MPEG, H.261, and H.263 ensure that software from different vendors will be compatible. Software which conforms to one of these standards is now readily available from several sources, e.g. IVS, VIC. The algorithm developed in this paper is compatible with H.261. While a custom coding scheme tailored specifically to sign language would offer better performance, the availability of a standard decoder is critical to ensure wide acceptance and availability. Sequences coded by this algorithm can be decoded by any H.261 coder, which offers many practical advantages.

This paper is organized as follows: Section 2 demonstrates the bandwidth requirements for sign language video. Section 3 introduces two Region-of-Interest schemes. Section 4 introduces the new algorithm and Section 5 presents results.

2 SIGN LANGUAGE BANDWIDTH

In the coding of video sequences, the available bandwidth must be apportioned between picture quality and frame rate, which are both important for sign language. Low frame rates require the user to move slowly, resulting in unnatural communication. The effect is similar to speaking slowly. The ability to convey personal feelings by varying articulation speed is lost. Poor resolution of face and hands would hinder finger spelling and speechreading. Significant compression is necessary to transmit quality images at a frame rate high enough for smooth motion perception over low bit rate channels. However, increasing the compression through computationally costly methods (such as motion compensation) will reduce the frame rate due to processing delays. Thus, a balance between algorithmic complexity and compression performance is required.

To demonstrate the increased motion of sign language sequences, tests were conducted to determine coding performance differences between a standard head-and-shoulders sequence and a sign-language sequence. Picture length, frame rate, and quantizer value were recorded during 20-second video sequences of both types, using a Silicon Graphics workstation (Indy) running IVS, a publicly available videoconferencing software package. IVS is an H.261 coder which has been adapted to the Internet. Complex packetization and error control schemes were developed to achieve reliable operation over packet switched networks [2].

Table 1 shows that the compressed sign language sequence had lower image quality: it had lower frame rate and lower resolution (higher quantizer value) compared to a head-and-shoulders sequence. The mean picture length was also larger,

Sequence	Band-width (kbps)	Frame Rate (fps)	Picture Length (bits)	FG Quantizer value
Head & Shoulder	64	18.0	162	12.5
Sign Language	64	15.4	285	19.0
Head & Shoulder	128	18.1	301	10.3
Sign Language	128	17.1	417	15.1

Table 1: Comparison between different 20 sec. QCIF sequences at two different bandwidth limits.

indicating the need for greater compression. If transmitted over a band-limited channel, then sign language video will require alternate coding strategies to achieve the same level of image quality.

3 REGIONS OF INTEREST

Coding standards such as H.261 are block-based. Most implementations of these standards give equal importance to each block. While different blocks within the same picture may be coded with different modes, no one block is more important than another. This model is not appropriate for sign language. Blocks which correspond to hands and face are more important than blocks in the image background. Allocating more bandwidth towards the quality of areas perceptually important to sign language comprehension, while sacrificing background quality, is a better coding strategy for sign language.

One way to accomplish this goal is to give priority to a ROI. For the same number of bits per picture, more bits would be allocated to pixels in the ROI, resulting in higher quality for the ROI and lower quality for the rest of the picture. The ROI may be defined statically or dynamically. A static ROI should be shaped to include the sign space, a region around the head and torso in which signs are performed [3] (Figure 1). A dynamic ROI should locate the face and hands in the image in real time and use the identified regions as the ROI.

Other authors have investigated video compression schemes that give priority to a ROI. Several proposals for MPEG-4 included general region-based coding schemes. Many earlier proposals use strictly the face as the ROI (although face segmentation methods vary widely) [4, 5, 6]. Moreover, many of these methods are incompatible with existing standards, and are computationally expensive; they cannot achieve high frame rates on current hardware. The methods in [7] and [8] are compatible with H.261. Once a ROI is defined, quantizer values are adjusted to give more bits to the ROI. Neither method took advantage of the ability to selectively encode macroblocks or attempted to control the frame rate. They also included only the face in the ROI.

A system for sign language must include the hands in the ROI. This paper presents a coding algorithm that used a static or dynamic ROI that included both

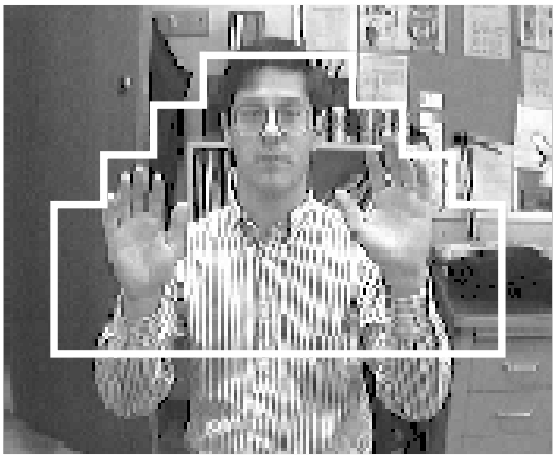

Figure 1: Static Region of Interest.

face and hands. The static ROI encompassed the sign space. The dynamic ROI tracked the speaker's face and hands. The developed algorithms were tested against a uniform coding scheme.

4 METHOD

Significant modifications were made to IVS to incorporate the two ROI coding methods introduced in Section 3. The only difference between the two methods is the way in which ROI macroblocks are defined. In the static method, certain macroblocks are designated as ROI *a priori* (Figure 1). A flesh-finding algorithm [9] is used to define ROI macroblocks in the dynamic method. Once the ROI macroblocks are defined, the two methods operate identically. These new methods were compared with a uniform method, in which all macroblocks are treated equally.

The developed coder was designed to sacrifice background image quality to maintain foreground quality and frame rate. There are three quantities which may be adjusted to avoid violating bandwidth constraints: (1) The quantizer value Q, which determines resolution; (2) The motion threshold T, which determines how much change is necessary between pictures for a macroblock to be sent; (3) The frame rate R. The uniform algorithm creates a table of (Q, T, R) 3-tuples, such that moving down/up rows in the table increases/decreases the compression rate. When the bit rate approaches the limit, the coder moves down one row and uses the (Q, T, R) values in the new row. Conversely, when the bit rate falls, the coder moves up a row, producing a higher quality picture. The ROI algorithms contain a table of (Q_F, Q_B, T_F, T_B, R) 5-tuples, where the F and B subscripts indicate foreground (ROI) and background (non-ROI), respectively. The only difference between the uniform and ROI methods is that different Q and T values are used in the ROI and non-ROI areas of the image.

Algorithm	Band-width (kbps)	Frame Rate (fps)	Picture Length (bits)	FG Quantizer value
Uniform	64	15.4	285	19.0
Static ROI	64	16.3	252	18.1
Dynamic ROI	64	16.3	237	17.6
Uniform	128	17.1	417	15.1
Static ROI	128	17.5	395	14.5
Dynamic ROI	128	17.8	379	13.7

Table 2: Mean values for grayscale QCIF pictures on SGI.

5 RESULTS

A grayscale QCIF video sequence was used to test the developed algorithm. The CIF size was not tested because the Silicon Graphics Indy workstation used for the tests was not able to encode a CIF sequence without encountering processing delays during periods of motion. Limiting the test to the QCIF size eliminated the computational load as as a possible cause of frame rate reduction. Therefore a drop in frame rate can be attributed to the coding algorithm. The sequence consisted of a person using sign language for 20 seconds. The person was positioned so that the signing occured within the static ROI. The (Q, T) (or (Q_F, Q_B, T_F, T_B)) values and picture length were recorded for each picture. The bit rate and frame rate were recorded once per second.

The test results for the three algorithms, which are shown in Table 2, confirm that the proposed algorithms address the requirements of sign language video. The results show that both the static and dynamic ROI algorithms give better compression, and slightly improve the frame rate and foreground quality compared to the uniform algorithm. The quantizer values for the non-ROI portions of the image, which are not shown in Table 2, are much higher indicating lower resolution.

Figure 2 shows the difference between the uniform and ROI schemes in terms of image quality. Both images show a snap shot of a moving head and shoulders sequence. The face in Figure 2(b) is sharper than in Figure 2(a), while the background is sharper in Figure 2(a).

ACKNOWLEDGMENTS

Funding for this project was provided by the National Science Foundation, grant #: HRD-9450019. Additional support has been provided by the Nemours Research Programs.

(a) Uniform (b) Static ROI

Figure 2: Moving head, coded with uniform and static ROI algorithms.

References

[1] U. Bellugi and S. Fischer, "A comparison of sign language and spoken language," *Cognition*, vol. 1, pp. 173–200, 1972.

[2] T. Turletti and C. Huitema, "Videoconferencing on the Internet," *IEEE/ACM Trans. Net.*, vol. 4, pp. 340–351, June 1996.

[3] E. S. Klima and U. Bellugi, *The Signs of Language*. Cambridge, Massachusetts: Harvard University Press, 1979.

[4] G. Bedini, L. Favalli, A. Marazzi, A. Mecocci, and C. Zanardi, "An integrated approach for high-compression of videoconference sequences," in *Proc. IEEE Intl. Conf. Comm. 95*, vol. 1, pp. 563–567, 1995.

[5] A. Eleftheriadis and A. Jacquin, "Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit rates," *Signal Processing:Image Communication*, vol. 7, no. 3, pp. 231–248, 1995.

[6] J. Luo, C. W. Chen, and K. J. Parker, "Face location in wavelet-based video compression for high perceptual quality videoconferencing," in *Proc. ICIP 95*, vol. 2, pp. 583–586, 1995.

[7] A. Eleftheriadis and A. Jacquin, "Low bit rate model-assisted H.261-compatible coding of video," in *Proc. ICIP 95*, vol. 2, pp. 418–421, 1995.

[8] E. Badiqué, "Knowledge-based facial area recognition and improved coding in a CCITT-compatible low-bitrate video-codec," in *Picture Coding Symposium*, 1990.

[9] R. Schumeyer and K. Barner, "Adaptive training of skin color distributions for segmentation of video sequences." Submitted to *ICIP 97*.