A VIDEO CODING SYSTEM FOR SIGN LANGUAGE COMMUNICATION AT LOW BIT RATES

Dimitris Agrafiotis[1], *Nishan Canagarajah*[1], *David R. Bull*[1], *Jim Kyle*[2], *Helen Seers*[2], *Matthew Dye*[3]

[1]Dept. of Electrical & Electronic Engineering, University of Bristol, UK
[2]Centre for Deaf Studies, University of Bristol, UK
[3]Presently with Dept. of Brain and Cognitive Sciences, University of Rochester, NY

ABSTRACT

The ability to communicate remotely through the use of video as promised by wireless networks and already practiced over fixed networks, is for deaf people as important as voice telephony is for hearing people. Sign languages are visual-spatial languages and as such demand good image quality for interaction and understanding. In this paper, based on analysis of the viewers perceptual behavior and the video content involved we propose a sign language video coding system using foveated processing, which can lead to bit rate savings without compromising the comprehension of the coded sequence. We support this claim with the results of an initial comprehension assessment trial of such coded sequences by deaf users.

1. INTRODUCTION

Remote communication trough the transmission of video over fixed or wireless networks is very important to the deaf community because it allows deaf people to communicate in their own language, sign language. Many video coding systems have focused on the compression of typical video conferencing sequences where a head and shoulders view of the participant is usually involved. Sign language video includes, in addition, the rapidly moving hands and arms of the imaged signer resulting in increased bit rate requirements[1]. This emphasizes further the need for efficient compression especially at low bit rates. This paper presents a video coding system for low bit rates adapted to the requirements of sign language (SL) communication. First an analysis of the needs and characteristics of SL video communication is presented. Based on this analysis the proposed system is then described. Coding results obtained with this system follow. Finally the results of an initial comprehension and quality assessment are given before concluding this work.

2. ANALYSIS OF SL VIDEO COMMUNICATION

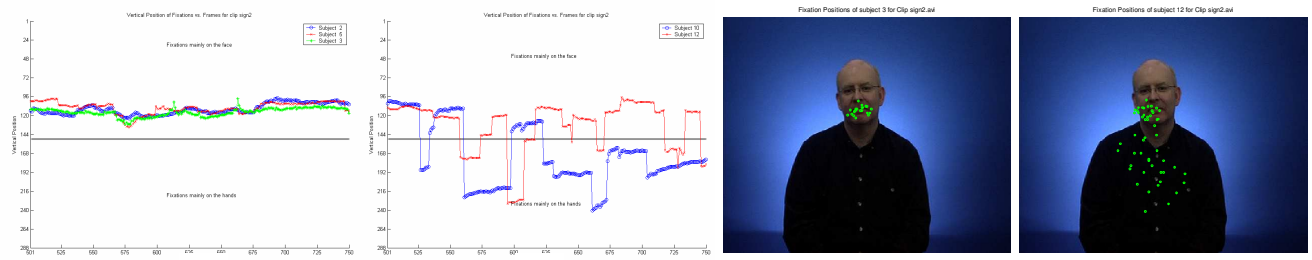
In order to propose a system for SL video communication it is necessary to first analyse the specific case with the aim of finding possible requirements and characteristics which should be fulfilled/exploited by such a system.

2.1. SL video viewers

In order to examine the SL video viewers behavior – how sign language viewers watch/perceive SL video material – a gaze-tracking study was setup, wherein the viewers eye-gaze was tracked while watching sign language video clips [2][3]. More specifically 28 subjects took part in experiments involving the use of an eye-tracking system[4] which recorded the participants eye-gaze while watching four one-minute clips showing short narratives being signed in British Sign Language(BSL) by an expert signer, sitting in front of a plain blue background. The clips were displayed uncompressed in the CIF format (352x288, 4:2:0) at 25 frames per second (fps). The participants included deaf and hearing BSL signers (i.e. interpreters) and hearing beginners/non-signers. Software written and used in the experiments reported results as fixation locations per frame (i.e. locus at which eye-gaze was directed), as well as overall eye movement events (which are either fixations or jumps/saccades).

Analysis of the results showed that sign language viewers, excluding the hearing beginners, concentrate on the face of the signer. In fact closer look at the results and considering the system's accuracy for the specific experimental setup (viewing distance, screen resolution, calibration) strongly suggests that SL viewers concentrate on the mouth of the viewed signer. There were no fixations on the hands except where the hands occlude the mouth. In contrast, naïve users of the language tend to follow the hands (mainly due to a lack of understanding), thus giving a much more spread viewing pattern. The results are graphically illustrated in figure 1 (left column), in terms of the vertical position of fixations per frame for 250 frames of clip 2 with respect to the clip's position on the screen. Two graphs are shown one with results of experienced SL viewers (a) and one of naïve users (b). The horizontal line represents a threshold for the position of the face (anything below that corresponds to fixations on the body or hands). The results are also visualised in terms of average fixation location for the duration of the clip (right column) for one experienced (c) and one naïve (d) viewer. Results are superimposed on the first frame of the specific test sequence.

0-7803-8554-3/04/$20.00 ©2004 IEEE.



a b c d

FIGURE-1 : Gaze tracking results of experienced (a,c) and naïve viewers (b,d) (see section 2.1 for description)

2.2. SL video content

Due to specific features of BSL, SL video content displays certain characteristics. As with the case of hearing person-to-person video conferencing, the signer at either end is the main point of focus and as such is located close to the centre of the viewing area. However in SL communication the hands (and hence the upper body) as well as the face and shoulders of the signer have to be visible, since they play a key role in the language, requiring an increased field of view for the cameras at each end. As a result a large part of the background is usually visible which not only is irrelevant for SL communication but can also be perceived as increased noise (distraction) in the viewers visual field. Apart from any motion in the background, the main activity in SL video consists of facial expression and head/hand changes. In conversation with another person, the signers position may also change but not usually to the extent of altering the body shape on screen – i.e. it may rotate but not move around the screen. These characteristics have an effect on the number of bits generated by each macroblock (MB) of the video frame when coded by a typical hybrid video coder (H.264 is used in this work). The amount of bits generated depends on the amount of activity in that particular MB, the effectiveness of the prediction and the quantization parameter (QP) used for coding the transform coefficients. A typical bit distribution of SL video is shown in figure 2 where the number of bits required to code each macroblock with QP=30 in 1 frame of a (CIF) plain-background sequence is depicted. Each square represents the location of one MB, and the brightness of the square specifies the number of bits spent. It can be seen that MBs corresponding to the position of the hands require a large number of bits.

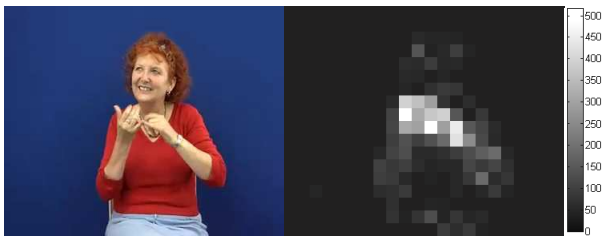


FIGURE-2 : Typical bit distribution of a coded SL video frame (QP30).

3. PROPOSED SL VIDEO CODING SYSTEM

The recorded viewing pattern confirms that the central point of fixation is the mouth and the rest of the moving image is seen with decreasing acuity. The fact that the hands apparently play an important role in the lexicon but are never fixated suggests that their motion and shape is processed only in peripheral vision. That is where the background is processed too (but probably discarded).

3.1. Foveated processing

A coding approach that follows this visual processing model is that of foveated video coding [5][6]. Foveated video compression aims to exploit the fall off in spatial resolution of the human visual system away from the point of fixation in order to reduce the bandwidth requirements of compressed video. We have followed the local bandwidth approach to foveated coding described in [6]. This method removes spatial high frequency components from regions away from the point of fixation which lowers the entropy of the video allowing increased coding gain. If the normalized viewing distance (normalized with regard to the physical size of the pixels on the screen) is known along with the point of fixation for every frame, then one can filter out a number of high frequencies without causing any perceived reduction in quality. For lossy coding we don't need to know the exact viewing distance. Instead we use it as a means of controlling the amount of loss. Hence the major obstacle is finding the fixation point, since this normally requires real time tracking of the viewers eye-gaze. The result of our gaze-tracking study removes this obstacle for the case of SL video since the fixation point will (almost) always lie on the face of the signer and close to the mouth. It introduces however the (simpler) need of locating the face of the displayed signer. The video image is partitioned into 8 regions based on their eccentricity (viewing angle) with the regions being constrained to be the union of disjoint MBs. More details can be found in [2][6]. Foveated processing produces a map showing the region each MB belongs to in each frame. In this work the foveation map is used to assign a different QP to each MB, with MBs lying in regions away from the fixation point being allocated a higher QP. The QP controls the amount of compression and corresponding fidelity reduction for each MB.

3.2. Face location/tracking

A large number of face detection/tracking methods exist, a good survey of which can be found in [7]. Methods can be classified into different categories based on the main approach used. We have employed a cascade of such methods combined with temporal information to track the signer's face in an SL sequence. Skin colour segmentation is first performed in the UV colour space, followed by the hierarchical multiscale approach of [8] applied to pixels classified as skin. The result of the previous steps is a number of face candidates based on the detection and relative arrangement of facial features. The face candidates are then passed on to a template matching module which verifies true faces based on their correlation with stored templates. If these are more than one then knowledge about the SL video is used to find the one that is most likely to be the signer's face. If the algorithm fails (most likely cause being the hands occluding facial features) then the results of the skin colour module together with temporal information are used to give the signer's face position. Apart from the first frame, the whole process is applied to the MBs corresponding to the signer's face in the previous frame along with a ring of MBs situated around them. This makes the whole method very robust and relatively fast.

3.3. Variable Quality H.264 coding

The foveation map described in the previous step is combined with a given range of QP values to produce MB regions that will be quantised with different step sizes, with the step size getting bigger for regions of higher eccentricity. An algorithm was written which ensures that outer regions always have their QP increased before inner regions and that the highest QP in the range is assigned to the lowest priority region for the 8 different foveation regions corresponding to the video frame shown figure 3. Region 0 is the highest priority region around the face the extent of which is given by the face tracking module. Coding with such a variable QP (VQP) incurs only a small overhead due to the coding of the difference in QP values (QP_{delta}) of MBs lying on region borders. A typical overhead for a CIF sized coded sequence (268 frames) with a QP range of 30-40 was found to be approximately 2.925 Kbits/sec (~3.5% of the actual rate). A block diagram of the final proposed system is shown in figure 4.

4. CODING RESULTS

Coding results are given for 2 clips, a plain background indoor scene, and an outdoor scene. The H.264 reference software (ver. 7.3) was modified in order to enable variable quantisation based on a given foveation map. The output bitstreams conform to the H.264 baseline profile. Five past reference frames were used for prediction. The input frame rate was 25fps, and the output 12.5 fps.

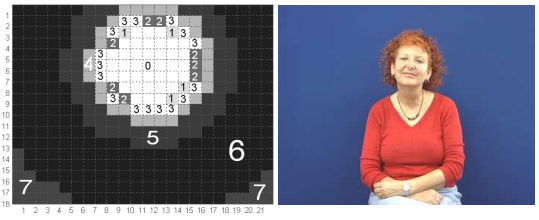


FIGURE-3 : QP allocation to foveation regions for a given QP range.

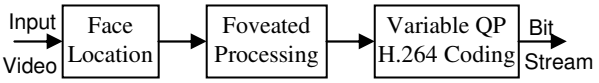


FIGURE-4 : Block diagram of proposed system.

Due to the lack of rate control, VQP coded sequences are compared in terms of resulting bit rate with Constant QP (CQP) ones, where the QP assigned to MBs of the high priority region (face) is the same for both versions (essentially similar quality for the face). It can be seen (table 1) that the proposed system leads to significant bit rate savings (~30%) compared to a standard CQP approach while keeping the quality of the important regions to similar high levels (figure 5). The rate savings are large partly because the region that generates the largest amounts of bits (the hands) undergoes coarser quantisation especially when located far way from the face. When closer to the face the hands are coded with higher fidelity (since they enter higher priority regions). This fits nicely with suggestions made in an early paper in the workings of sign language [9] according to which the language has/is evolving to accommodate a face centred viewing pattern by bringing the hands closer to the face when detailed signs have to be made, leaving mostly gross movements and gestures to take place in regions away from it.

TABLE 1: Constant QP vs. Variable QP results.

Sequence	CQP 30	VQP 30-40	Bit Rate Reduction
Indoor	112.84 Kbps	74.94 Kbps	33 %
Outdoor	130.22 Kbps	90.87 Kbps	30 %

FIGURE-5 : Coding results with the proposed method.

As expected, the use of larger QP values for MBs lying in peripheral regions leads to increased fidelity reduction for the low priority regions especially after long occlusions e.g. in an outdoor scene after a car passes by in the background (assuming the max number of reference frames is 5). Due to the in-loop filter used in H.264 this demonstrates itself as increased blurriness. To moderate this effect one can keep the first I-frame of the sequence (IDR frame) – which is coded with a constant QP equal to QPmin – in memory as a long term reference frame. Because the camera is generally fixed, there are many regions in the background that do not change significantly and which can be predicted more efficiently from this first frame after such occlusions. Making the first frame of a sequence a long-term reference frame requires only raising a flag in the slice header. The resulting bit-rate is similar but the subjective and objective quality (PSNR) of the lower priority regions is improved for sequences with increased background activity. The quality of the high priority region remains roughly the same.

TABLE 2: High and Low priority region PSNR.

Sequence		Region 0	Regions 1-7	Total Bit rate
Indoor	Short Term IDR	34.91 dB	35.81 dB	74.94 Kbps
	Long Term IDR	34.95 dB	35.90 dB	75.48 Kbps
Outdoor	Short Term IDR	32.92 dB	30.25 dB	90.87 Kbps
	Long Term IDR	33.02 dB	31.70 dB	90.21 Kbps

FIGURE-6 : Detail of decoded frame with a short term IDR frame (left) and a long term IDR (middle). The original frame is also shown (right).

5. COMPREHENSION ASSESMENT

In order to assess the effect of the proposed coding approach on the ability to comprehend the coded SL video a small trial was setup with 17 deaf participants who watched 2 clips with and without background activity respectively. Each clip was separated in 3 segments with each segment being assigned and coded randomly with a QP range of 30 (i.e. CQP), 30-36 and 30-40. The trial aimed to assess the perceived quality while watching and understanding (at the same time) the content of the clips. The participants were asked to rate the clips in terms of quality and blurriness on a scale from 0 to 100. The level of comprehension was also assessed by asking questions related to the signed content after watching each segment. The comprehension results were at a ceiling indicating that the proposed coding approach does not affect understanding even though it introduces losses in quality at peripheral regions. In terms of perceived quality the plain background sequence received a similar rating for all cases while the outside sequence received a lower rating for the 30-40 version. Differences in blurriness were not perceived as much with the 30-36 and 30-40 plain background sequences as with the corresponding outside ones. The sequences used in the trial did not employ a long-term IDR frame.

6. CONCLUSION

Coding of image sequences will always result in some information being lost in order to satisfy rate requirements set by the network over which transmission will take place. In this paper we have proposed and described a sign language video coding system with which it is possible to localize this information loss, in a way that should not impair sign language comprehension. The system employs variable quality coding based on foveated processing of the input video frames which requires tracking of the imaged signer's face in the clip. The results presented are very promising, indicating that substantial bit rate savings can be had without affecting the comprehension ability of the viewers.

AKNOWLEDGEMENTS

The authors would like to thank the DTI - UK for funding the project.

REFERENCES

[1] R. P. Shumeyer, E.A. Heredia, K.E.Barner, "Region of Interest Priority Coding for Sign Language Videoconferencing", IEEE Workshop on Multimedia Signal Processing, pp. 531 -536, 1997.

[2] D. Agrafiotis, N. Canagarajah, D.R. Bull, M. Dye, H. Twyford, J. Kyle, J. Chung-How, "Optimised Sign Language Video Coding Based On Eye-Tracking Analysis", SPIE Int. Conf. on Visual Communications and Image Processing (VCIP), Lugano, Switzerland, July 2003.

[3] D. Agrafiotis, N. Canagarajah, D.R. Bull, M. Dye " Perceptually Optimised Sign Language Video Coding Based on Eye Tracking Analysis", IEE Electronics Letters, vol. 39, no.24, Nov. 2003, pp. 1703-1705.

[4] SR Research, "The Eyelink System", http://www.eyelinkinfo.com/.

[5] Wilson S. Geisler, Jeffrey S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication", SPIE Proceedings, vol. 3299, 1998.

[6] H.R. Sheikh, S.Liu, B.L.Evans, A.C.Bovic, "Real Time Foveation Techniques for H.263 Video Encoding in Software", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, vol.3, pp. 1781-1784, 2001.

[7] Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, "Detecting Faces in Images: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, January 2002.

[8] Jun Miao, Baocai Yin, Kongqiao Wang, Lansun Shen, Xuecun Chen, " A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template", Pattern Recognition, vol 32 (1999) pp. 1237-1248.

[9] P. Siple, "Visual constraints for sign language communication", Sign Language Studies, 19, pp. 95-110, 1978.