Gaze tracking and its application to video coding for sign language

Laura Muir, Iain Richardson and Steven Leaper

Image Communication Technology Group, The Robert Gordon University, Schoolhill, Aberdeen, UK
l.muir@rgu.ac.uk, i.g.richardson@rgu.ac.uk

ABSTRACT

Sign language communication via videotelephone has demanding visual quality requirements. In order to optimise video coding for sign language it is necessary to quantify the importance of areas of the video scene. Eye movements of deaf users are tracked whilst watching a sign language video sequence. The results indicate that the gaze tends to concentrate on the face region with occasional excursions (saccades). The implications of these results for prioritised coding of sign language video sequences are discussed.

1. Introduction

Deaf people communicate using the media of sign language and lip reading. Sign language communication is based on hand shapes and movements, supported by finger spelling, lip movements, facial expression, eye movements and body language. The rich visual structure of sign language makes it possible to communicate complex concepts rapidly and accurately.

The advent of visual telecommunications is seen as very important to the deaf community because of the potential for sign language communication at a distance, offering greater freedom than text-based alternatives [1]. However, the shortcomings of existing videoconferencing and videotelephony systems can be frustrating for sign language users.

Sign language contains rapid, detailed hand movements that convey concepts and spell out words. At the same time, an experienced signer articulates the shape of the words with her mouth and conveys information through facial expression and body language. Figure 1 shows 2 frames from a video clip of British Sign Language (BSL). ITU-T draft profile [2] proposes quality requirements for sign language communication via videoconferencing systems. This profile specifies CIF resolution (352x288 luminance samples per frame) and a frame rate of 25 frames per second as a minimum for accurate sign language communication.

Current videotelephony standards such as H.263 / H.263+ [3] can provide reasonable visual quality and frame rates if a high bitrate connection is available, but at lower bitrates (under c. 200kbps), video is characterised by low frame rates, small picture sizes (QCIF or less) and/or poor decoded quality. Even with the improved compression efficiency provided by H.264 [4], video quality over low bitrate channels may fall short of what is required for effective sign language communication. Deaf users describe the problems of having to slow down and exaggerate hand movements to cope with poor image quality and low frame rates; this makes sign language communication via videophone tiring and limits its usefulness to the deaf community.

Figure 1 Sample frames from sign language sequence

It may be possible to improve the quality of sign language communication by selectively prioritising areas or components of the visual scene based on features and/or motion [5,6]. A proposal for priority encoding of sign language video is described in [7] but does not address temporal and spatial quality requirements for effective sign language communication, nor does this paper report results of testing with sign language users. It is necessary to identify and quantify the importance of components of a sign language video scene in order to develop effective prioritisation

algorithms. The gaze direction of an experienced sign language user during a sign language conversation may give important clues about the relative importance of areas of the scene. This paper describes a set of experiments to record and analyse the gaze direction of sign language users whilst watching a video sequence of BSL.

The experimental equipment and method are described in section 2; section 3 presents results of the experiments; section 4 discusses the results and their application to prioritised video coding.

2. Method

An experiment was carried out with a group of 8 deaf users, categorised as "L1" (deaf from birth, with BSL as their first language). During the experiment, the gaze direction of each user was recorded whilst watching a video clip of BSL.

Two frames from the test video clip are shown in Figure 1. The clip lasts for 1 minute and was played twice for each user (with a series of calibration markers between the two clips). The sequence had not been seen by any of the subjects before the experiment.

The video clip was played full screen on a PC monitor and each participants watched the clip from a fixed viewing distance. Eye movements were recorded during playback using a "Quick Glance" eye tracking system. This system uses infrared illumination of the subject's pupil to record X-Y gaze point coordinates at a rate of 30Hz.

The gaze direction coordinates were saved to a text file and the sections of the file corresponding to the two 1-minute video clips were extracted for analysis.

3. Results

3.1 Eye movement results : spatial

The set of (x,y) gaze coordinates for 3 of the test subjects (User A, User B and User C, all L1 deaf subjects) are plotted in Figure 2. These plots correspond to the 2nd play-through of the video clip; similar results were obtained from the 1st play-through. The eye movements of all 8 subjects exhibit significant similarities. In each case, the gaze is concentrated on the face of the signer in the video clip, with occasional "excursions" to other regions mostly in a vertical axis through the face of the signer (see plots for User A and User C). Two of the subjects (one of which was User B, shown here) showed very precise concentration with few excursions.

The results for Users A, B and C are presented as 2-D histograms in Figure 3. These histograms plot the number of gaze points occurring in each 10x10-pixel square in the image. The majority of the points occur around the face region of the signer ("bright" area) with a small number of excursion points outside this region.

Figure 4 shows the scatter plot for User B, overlaid with circles representing viewing angles relative to a central point. The circles

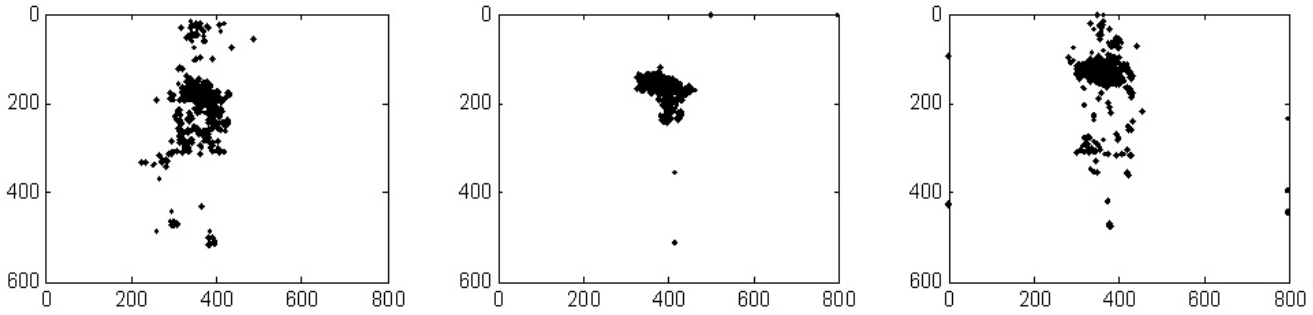

Figure 2 Scatter plots: User A, User B, User C

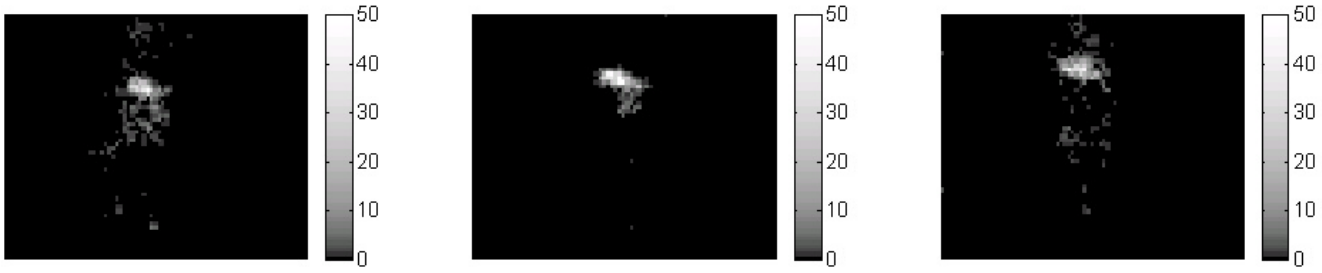

Figure 3 Histograms: User A, User B, User C

are centred on the median position of all the samples (the approximate centre of the subject's gaze during the sequence) and are plotted at constant angles of 2.5°, 5° and 10° from the centre. It is clear from this Figure that most of the gaze points lie within 2.5° of the centre: over 75% of points fall within this circle for Users A and C and over 90% for User B.

3.2 Eye movement results: temporal

Figure 5 plots the Y-coordinate of User A's gaze against time for the second play-through of the video clip. The median Y-coordinate is 185 and the plot clearly shows that the gaze is concentrated mostly around this position with occasional excursions. Excursions of the subjects' gaze away from the median typically last for less than 0.5 seconds and are concentrated in the vertical axis through the centre point. In the video sequence, the signer looks down and turns a page of her notes at around 95 seconds; this corresponds to a large downward excursion shown in the Figure. There is a similar significant excursion at the same point in the other sets of results. The remaining excursions are less pronounced.

3.3 Residual coefficient analysis

The test video clip was encoded using an H.263 Baseline encoder (with a fixed quantizer step size of 8) and the number of non-zero coefficients remaining in each macroblock after motion compensation, DCT and quantization were counted. Figure 6 plots the total number of non-zero quantized residual coefficients in each macroblock position for the first 200 inter-coded frames of the sequence.

This Figure shows that the non-zero coefficients are concentrated around the head, upper body and arm regions of the signer. This is as might be expected since the camera position is fixed and the signer does not move her standing position significantly during the sequence. However, it is interesting to note that the highest concentration of non-zero coefficients occurs in the lower-left and lower-right regions of the scene. The signer moves her hands through a number of positions (as illustrated by Figure 1) but the most significant residual energy due to hand movements appears in these two regions. The residual coefficient energy in the face region is relatively low in comparison.

4. Discussion

4.1 Eye tracking

The eye movement tracking results indicate a response to the experiment that is reasonably consistent across all subjects. In each case, the gaze is centred on a point roughly in the middle of the face of the signer. We believe that this is due to the fact that facial expression, lip shape and eye position convey important

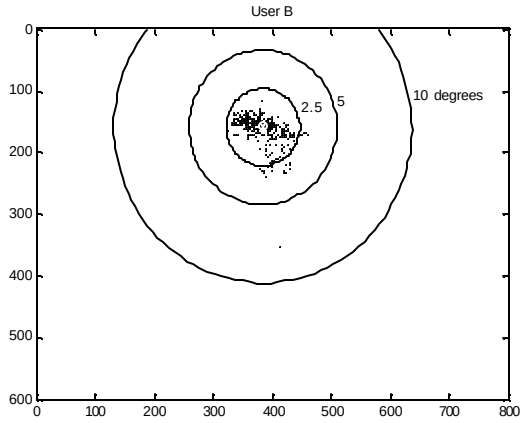

Figure 4 User B: scatter plot showing angular distribution

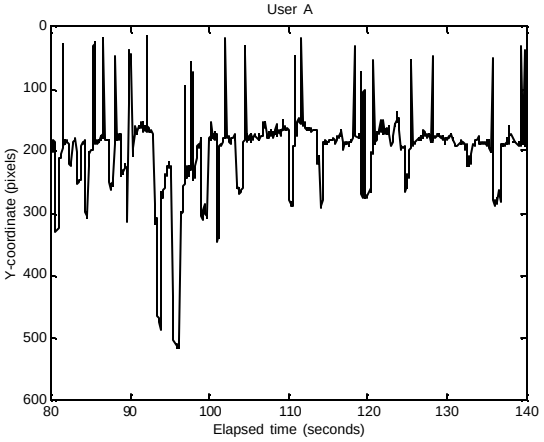

Figure 5 User A: plot of Y coordinate vs. time

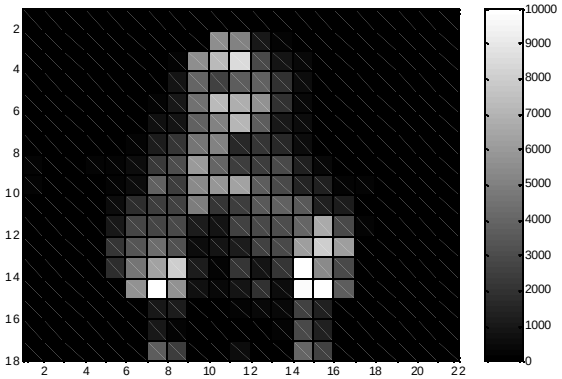

Figure 6 Count of non-zero quantized coefficients in each macroblock position (QP=8, first 200 frames of sequence)

information during a sign language conversation and these features may require closer visual attention than the more expansive movements of hands. Most of the recorded gaze points occur within a viewing angle of 2.5^o relative to the centre point. The effective resolution of human vision drops by a factor of two at a viewing angle of 2.5^o from the point of attention and reduces logarithmically beyond this angle. Six of the subjects' responses included frequent excursions from this central area, usually along a vertical access, lasting up to 0.5 seconds but typically around 150-200ms. During a short excursion (or "saccade"), high-resolution vision is suppressed by the Human Visual System (saccadic suppression) [8]. These results imply that an experienced sign language user perceives the face region with high visual resolution throughout a sign language conversation, in order to extract information from the face, lips and eyes. Hand and body movements are perceived mainly in lower-resolution peripheral vision. Occasional saccadic excursions away from the face area are made (by some subjects), possibly to view specific hand movements with increased accuracy. However, these excursions are usually too short to "see" the hand / body region in full spatial detail.

4.2 Application to video coding

These results have implications for the coding of sign language video sequences. The results imply that the face region should be rendered with the highest possible spatial and temporal fidelity. Hand and arm movements are largely seen in lower-resolution "peripheral" vision and therefore may not require such high spatial fidelity. However, sign language users pick up important information due to rapid movements of the hands and arms [2] and so we would expect these regions to require high temporal fidelity.

A suitable strategy for coding sign language video may therefore be to prioritise certain regions of the image in order to provide high spatial quality and temporal resolution to the face region and high temporal resolution to the arms / hands. Anectodal evidence from experienced signers indicates that the background region is unimportant (and in fact a detailed or "busy" background is distracting) which implies that this region may be rendered at a low spatial quality and temporal resolution without detriment to successful communication.

Options for prioritisation include object-based coding, macroblock-level prioritisation and pre-processing. Object-based coding (for example using the tools provided by MPEG-4 Visual Main and Core Profiles) gives perhaps the highest flexibility in coding foreground and background objects with different coding parameters and refresh rates. However, accurate segmentation along object boundaries is computationally intensive and there is a lack of practical implementations of the object-based tools within MPEG-4 Visual. A more practical approach may be to selectively prioritise macroblocks during coding with a block-based scheme such as MPEG-4 Visual Simple Profile or H.264. Controlling quantization parameter and macroblock skip mode makes it possible to provide varying spatial quality and temporal update rate in different regions of the scene. Segmentation need only be accurate to within a 16x16 macroblock boundary. However, this approach can only support prioritisation of very approximate regions. Pre-processing of the video image to remove spatial detail prior to encoding is an alternative approach to selective coding. For example, foveated processing renders an image at reduced resolution as the distance from a priority region ("fovea") increases, mimics the processing of the human visual system. Reducing the resolution of outlying regions in this way has the potential to significantly reduce the bitrate of compressed video [9]. It may be appropriate to prioritise both the face and the hands in this type of scheme.

Based on the results of encoding the test sequence using H.263 (section 3.3), it is clear that a relatively small proportion of residual coefficient energy is concentrated in the visually important face region when a uniform quantization parameter is applied. This indicates that prioritising the coded data (and hence allocating more coded bits to selected regions) has significant potential for improving the quality of the visually important regions.

5. Conclusions

The results of this experiment demonstrate that experienced sign language users exhibit a consistent, characteristic eye movement response whilst watching sign language. There is potential to exploit this response in order to improve the subjective quality of coded sign language video sequences. The user's attention is focussed on a well-defined region around the signer's face with occasional saccades to the region of the hands. This implies that it may be possible to prioritise regions of the image spatially and/or temporally. By adopting a prioritisation scheme and optimising spatial quality around the face region and temporal quality around the face and arms/hands, it should be possible to make significantly better use of the available bitrate and improve the subjective quality of sign language communication.

Further work is required to implement and evaluate a prioritised coding scheme for sign language video. We believe that it is also necessary to develop new metrics for subjective quality of sign language. Objective metrics such as PSNR and subjective measures such as those defined in ITU-R Rec. BT.500 [10] are not necessarily appropriate because the important issues are clarity and ease of sign language communication (rather than measured quality across the entire video image).

6. Acknowledgements

The authors wish to record their sincere thanks to Lilian Lawson of the Scottish Council on Deafness for providing video material and advice and for Jim Hunter and other members of the Aberdeen

and North East Deaf Society for participating in the data collection experiments.

7. References

[1] McCaul, T F, Video-based telecommunications technology and the deaf community, Australian Communication Exchange Report, 1997.

[2] ITU-T SG16, Draft Application Profile: Sign language and lip-reading real time conversation usage of low bit rate video communication, September 1998.

[3] ITU-T Rec. H.263, Video coding for low bit rate communication, February 1998.

[4] DRAFT ISO/IEC 14496-10 : 2002 (E), Rec. H.264 , Advanced Video Coding, Geneva, October 2002.

[5] Eleftheriadis, A and Jacquin, A, Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit rates, Signal Processing: Image Communication, 7 (3), 1995.

[6] Favalli, L, Mecocci, A and Moschetti, F, Object tracking for retrieval applications in MPEG-2, IEEE Trans. Circuits and Systems for Video Technology, 10(3), 2000.

[7] Schumeyer, R, Heredia, E and Barner, K, Region of interest priority coding for sign language videoconferencing, IEEE Workshop on Multimedia Signal Processing, Princeton, June 1997.

[8] Delgado-García, J M, Why move the eyes if we can move the head ?, Brain Research Bulletin, Vol. 52 no. 6, pp 475-482, 2000.

[9] Geisler, W S and Perry, J S, A real-time foveated multiresolution system for low-bandwidth video communication. SPIE Proceedings Vol. 3299, 1998.

[10] ITU-R Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures, 2002.