VISUAL PERCEPTION OF CONTENT-PRIORITISED SIGN LANGUAGE VIDEO QUALITY

L. J. Muir, I. E. G. Richardson and K. Hamilton

Image Communication Technology Group,
The Robert Gordon University, Schoolhill, Aberdeen, UK.
Telephone 01224 262400; Fax 01224 262444
l.muir@rgu.ac.uk; i.g.richardson@rgu.ac.uk

Keywords: Visual Perception, Video Communication, Sign Language, Content-Prioritised Video Coding.

Abstract

Video communication systems currently provide poor quality and performance for deaf people using sign language, particularly at low bit rates. Our previous work, involving eye movement tracking experiments and analysis of visual attention mechanisms for sign language, demonstrated a consistent characteristic response which could be exploited to enable optimisation of video coding systems performance by prioritising content for deaf users. This paper describes an experiment designed to test the perceived quality of selectively prioritised video for sign language communication. A series of selectively degraded video clips was shown to individual deaf viewers. Participants subjectively rated the quality of the modified video on a Degradation Category Rating (DCR) scale adapted for sign language users. The results demonstrate the potential to develop content-prioritised coding schemes, based on viewing behaviour, which can reduce bandwidth requirements and provide best quality for the needs of the user. We propose selective quantisation to reduce compression in visually important regions of video images, which require spatial detail for small slow motion detection, and increased compression of regions regarded in peripheral vision where large rapid movements occur in sign language communication.

1 Introduction

Video compression research, development and standardisation have enabled the development of new visual communication applications which aim to bridge the gap between the requirements of the user and the limited capabilities of communication networks. The demand for video communication over networks is high but the performance of video telephony and video conferencing systems in particular has not met the quality and reliability standards which users require for it to be widely used for inter-personal communications. A user group which relies heavily on communication of visual information is the deaf community.

British Sign Language (BSL) is the first language of up to 70,000 deaf people in the United Kingdom (for whom English is a second language). Sign language is a rich combination of visual signals including facial expression, mouth/lip shapes, hand gestures, body movements and finger-spelling. Communication of visual information between deaf people during a freely expressed sign language conversation is detailed and rapid. Accurate personal communication of sign language at a distance places specific demands on a visual media application in terms of quality, speed, reliability and economy. These demands are not adequately met by current video communication solutions.

The minimum quality requirements for sign language video communication are CIF resolution (352 x 288 displayed pixels) and a frame rate of at least 25 frames per second [7]. At high bit rates, reasonable picture quality and frame rates can be achieved using the H.263 video coding standard [5]. At bit rates below 200 kilobits per second (kbps), real time communication of video is characterised by low frame rates, small picture size and poor picture quality [13]. Even the improved video compression efficiency of the new H.264 standard [4] may not be acceptable for accurate sign language communication at low bit rates. Deaf people have to modify their sign language, for example by using slow exaggerated movements, to overcome the limitations of videophones and this restricts the usefulness of video technology for the deaf community.

Previous work has investigated the efficiency savings which can be achieved using video content prioritisation schemes [2, 15]. Saxe and Foulds [14] proposed an image segmentation and region-of-interest coding scheme based on skin detection. Other work has used foveated processing to mimic human visual processing [3, 16]. Geisler and Perry [3] obtained bandwidth savings by matching the spatial resolution of transmitted images to the smooth decline in spatial resolution of the human visual system. This method exploited the properties of foveated vision and resulted in the development of a foveated multi-resolution pyramid video coder/decoder. The foveated regions were determined using a contrast threshold formula based on human contrast sensitivity data measured as a function of spatial frequency and retinal eccentricity. The compressed video in these methods

demonstrated efficiency gains but made assumptions about how the video material was viewed and the resulting video output was not subject to quality testing by the target end user.

Agrafiotis et al [1] propose a coding scheme which combines skin colour segmentation (to locate the face of the signer), foveated processing and variable quantisation in eight macroblock regions of the image. The authors demonstrate coding gains but there is no rationale for the eight-region model of foveation, indication of the complexity of the scheme or detail of the method and results of subjective testing.

Our eye movement tracking experiments [9, 10, 11] established that sign language users exhibit a constant characteristic eye movement response to sign language video. We found that a deaf viewer fixates mostly on the facial region of the signer in the video to pick up small detailed movements, associated with facial expression and lip/mouth shapes, which are known to be important for comprehension of sign language. Eye movements direct the fovea of the eye (which is responsible for high resolution vision) to the fixation point. The face is therefore seen in high spatial detail. Assuming that hand gestures play a significant part in sign language communication, it must be the case that they are observed in peripheral vision when they are not close enough to the face to be captured by the fovea of the eye. Peripheral, low resolution, vision was found to be adequate for gross and rapid sign language gestures that occurred away from the face region of the signer in our experiments. These findings support the theory, presented by Siple [16] that since sign language is received and processed initially by the visual system then the rules for communicating signs would be constrained by the limits of that system.

This paper describes an experiment, conducted with profoundly deaf volunteers, to test the perception of quality of video which had been modified based on the findings of our previous eye movement tracking experiments and the properties of foveated vision. The experimental method including the subjective quality assessment (based on [6]) is described in section two. Results are presented in section three and the findings and further work are discussed in section four.

2 Method

2.1 Subjects

Subjective quality assessment experiments were conducted with six profoundly deaf-from-birth volunteers. British Sign Language (BSL) is the first language and English the second language of all the subjects who participated in the experiment. For this reason communications were in BSL, aided by an interpreter who was known to the participants. The subjects had normal visual acuity or corrected-to-normal acuity.

2.2 Materials and Apparatus

The sign language video material for the experiment was captured at 25 frames per second on a SonyVX200E Digital Video camera, under controlled artificial lighting in the University video recording studio, using one profoundly deaf volunteer. The volunteer who signed in the video material is from the same geographical area, the North-East of Scotland, and used the same version of BSL (which has regional variations analogous to speech dialects) as the subjects participating in the experiment. The signer used facial expression, lip movement, gestures, detailed finger-spelling and body movement around the scene which had a plain background. She related short stories from her own experience using her own natural style and expression of signing. Short video clips were selected to ensure the test material contained a wide range of sign language movements, expressions and gestures (including finger spelling). In addition, five different clips were created for training the participants before the main experiment began.

The video clips were degraded using a modified version of the implementation of the Geisler & Perry foveation algorithm [3] developed by William Overall at Stanford University [8]. The clips were pre-processed by stepping through each one, frame-by-frame, marking the central point for foveation (in this case the tip of the nose of the signer) and degrading the spatial quality from that central point according to the minimum Contrast Threshold (CT_0) and viewing distance set for each video clip. The CT_0 and thus the degree of foveation blurring ranged from zero (none) to 0.2 (high). The other parameters set in the foveation algorithm remained constant; viewing distance = 0.305, spatial frequency decay constant (alpha) = 0.106 and half-resolution retinal eccentricity (e_2) = 2.3. The output of the foveation algorithm is a video clip which has a smooth reduction in spatial resolution from the point of foveation. An additional clip with enhanced foveation (clip 10) was created, using clip 3, by increasing the value of alpha to 0.212 and decreasing the value of e_2 to 1.15 (0.0156en in table 3).

Clip Number	Clip Name (Duration)	Contrast Threshold (CT_0)	English translation of BSL content
1	Bus Stop (10.20 seconds)	0.00	Standing at the bus stop, people could see I was deaf. They could see my hearing dog's jacket but they still talked. I didn't know what they were talking about.
2	Family (10.16 seconds)	0.0156	I have one brother and one sister. I am older than them. My brother is divorced.
3 (& 10)	Introduction (10.18 seconds)	0.03 (& 0.0156en)	Hello, my name is Lisa. My dog's name is Bran. He is a hearing dog for the deaf. He helps me.
4	Hobbies (7.22 seconds)	0.05	My hobbies, I love cooking, swimming and tap dancing.
5	Holiday (10.10 seconds)	0.075	I went on holiday to Spain last year. I had a good time. The weather was very warm.
6	School (8.03 seconds)	0.10	I went to Aberdeen School for the Deaf. As I was growing up I used signs and learned oral communication.
7	Worlds (7.02 seconds)	0.13	When I was at school there was a hearing world and a deaf world. Now I have both worlds.
8	Deaf (9.05 seconds)	0.16	When you meet a deaf person you think deaf people are the same. They are not; there are different levels of deafness.
9	Television (9.01 seconds)	0.20	When I watch TV, I look at the subtitles or the signer at the bottom right and I look back and to the TV picture.

Table 1: Video Clips for Subjective Testing

The video clips for the training session, conducted before the main experiment, were created with CT_0 values of 0.0156, 0.05, 0.075, 0.10 and 0.20.

The video clips were displayed to the viewer on a seventeen inch monitor with true colour, 32 bit display connected to a Dell Pentium IV PC with PCI Video Capture Card installed.

2.3 Procedure

The subjective quality assessment method used in the experiment was the Degradation Category Rating (Double Stimulus Impairment) method adapted from [6] for sign language users. The experiment was conducted with individual subjects positioned at a comfortable viewing distance from the PC monitor. Video clips were presented full screen one at a time in pairs. The first video in the pair was the source reference video clip and the second clip was the source video which had been degraded according to the CT_0 value set in the foveation algorithm. A plain mid-grey coloured screen was presented for two seconds between each clip in the pair and for ten seconds between each pair of clips. The subjects were asked to rate the quality of the second (foveated) video clip compared to the first (source reference) clip during the ten-second period (voting time) between pairs of clips. The five-point rating scale, with English translations of the descriptions adapted for sign language users, is given in table 2.

Rating	Description
5	Imperceptible difference
4	Perceptible difference but not annoying, the sign language was clear
3	Sign language is slightly unclear, one or two signs were not clear but the story was understood
2	Annoying, sign language was not clear making it difficult to understand the story
1	Very annoying, sign language was obscured and the story could not be understood

Table 2: Five-Point Degradation Category Rating Scale

The rating for each foveated clip was conveyed to the researcher in sign language. Prior to the main experiment, a training session consisting of six video pairs was conducted with each participant. The purpose of the training session was to familiarise the subject with the procedure and rating scale. The results for the training session were recorded but not used in the analysis of the findings. The main experiment consisted of six sets of five video pairs with short rest breaks between each set of video pairs. The clips were presented in random order (not according to the degree of foveation blurring) and each video was included three times in a different order during the experiment. The purpose of the

repetition was to allow reliability of scoring by individual subjects to be checked.

3 Results

Sign language conversations with the subjects after the experiment demonstrated understanding of, and interest in, the content of the video clips used. The subjective scores given by the six subjects during six sets of five video clip pairs (tests a to f) are given in Table 3. This table presents the ratings (described in Table 2) awarded by the subjects in each test. The Mean Opinion Score (MOS) and Standard Deviation (SD) are given for each viewing of the clips in the experiment. The subject's ratings, recorded at each of three instances of viewing the clips, and the MOS and Standard Deviation for each level of foveation are given in Table 4. The average MOS for each clip is illustrated in Figure 1.

a) Test 1		DCR Score per Subject								**d) Test 4**		DCR Score per Subject							
Video	Foveation	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD	Video	Foveation	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
10	0.0156en	3	5	5	5	3	5	4.3	1.0	5	0.0750	5	5	4	5	5	5	4.8	0.4
2	0.0156	5	5	5	5	5	5	5.0	0.0	8	0.1600	3	5	4	5	5	5	4.5	0.8
4	0.0500	4	5	5	5	5	5	4.8	0.4	6	0.1000	5	5	5	5	5	5	5.0	0.0
5	0.0750	5	3	5	5	5	5	4.7	0.8	2	0.0156	5	4	4	5	5	5	4.7	0.5
6	0.10	5	5	5	5	4	5	4.8	0.4	1	0.0000	4	5	4	5	5	5	4.7	0.5
b) Test 2		DCR Score per Subject								**e) Test 5**		DCR Score per Subject							
Video	Foveation	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD	Video	Foveation	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
1	0.0000	5	4	5	5	5	5	4.8	0.4	9	0.2000	5	4	4	3	5	5	4.3	0.8
3	0.0300	5	5	5	5	5	5	5.0	0.0	10	0.0156en	2	5	4	3	4	5	3.8	1.2
9	0.2000	3	5	5	4	4	5	4.3	0.8	5	0.0750	3	5	5	5	5	4	4.5	0.8
7	0.1300	5	5	5	5	5	5	5.0	0.0	4	0.0500	5	5	5	5	5	5	5.0	0.0
8	0.1600	2	5	5	5	5	5	4.5	1.2	8	0.1600	5	5	5	5	5	5	5.0	0.0
c) Test 3		DCR Score per Subject								**f) Test 6**		DCR Score per Subject							
Video	Foveation	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD	Video	Foveation	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
3	0.0300	5	5	5	5	5	5	5.0	0.0	7	0.1300	5	5	5	5	4	5	4.8	0.4
1	0.0000	4	4	5	4	5	4	4.3	0.5	9	0.3000	5	5	5	5	4	5	4.8	0.4
2	0.0156	5	5	5	5	5	4	4.8	0.4	6	0.1000	5	5	5	5	5	5	5.0	0.0
4	0.0500	3	5	4	5	5	5	4.5	0.8	3	0.0300	5	5	5	5	5	5	5.0	0.0
7	0.1300	4	5	4	5	5	5	4.7	0.5	10	0.0156en	5	5	4	5	4	5	4.7	0.5

Table 3: Subjective Quality Scores (Raw Data)

	Subject 1					**Subject 2**					**Subject 3**					**Subject 4**					**Subject 5**					**Subject 6**					**Average**	
Foveation	**1**	**2**	**3**	**MOS**	**SD**	**1**	**2**	**3**	**MOS**	**SD**	**1**	**2**	**3**	**MOS**	**SD**	**1**	**2**	**3**	**MOS**	**SD**	**1**	**2**	**3**	**MOS**	**SD**	**1**	**2**	**3**	**MOS**	**SD**	**MOS**	**SD**
0.0000	5	4	4	4.3	0.6	4	4	5	4.3	0.6	5	5	4	4.7	0.6	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	4	5	4.7	0.6	4.6	0.5
0.0156	5	5	5	5.0	0.0	5	5	4	4.7	0.6	5	5	4	4.7	0.6	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	4	5	4.7	0.6	4.8	0.4
0.0300	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	4	4.7	0.6	5	5	5	5.0	0.0	4.9	0.2
0.0500	**4**	**3**	**5**	4.0	1.0	5	5	5	5.0	0.0	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	4.8	0.5
0.0750	**5**	**5**	**3**	4.3	1.2	**3**	**5**	**5**	4.3	1.2	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	4	4.7	0.6	4.7	0.7
0.1000	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	4	5	5	4.7	0.6	5	5	5	5.0	0.0	4.9	0.2
0.1300	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	5	4	4.7	0.6	5	5	5	5.0	0.0	4.8	0.4
0.1600	**2**	**3**	**5**	3.3	1.5	**2**	**4**	**3**	3.0	1.0	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	4.3	1.1
0.2000	**3**	**5**	**4**	4.0	1.0	5	4	5	4.5	0.5	5	4	5	4.5	0.5	**4**	**3**	**4**	3.5	0.5	4	5	5	4.5	0.5	5	5	5	5.0	0.0	4.3	0.7
0.0156en	**3**	**2**	**5**	3.3	1.5	5	5	5	5.0	0.0	5	4	4	4.3	0.6	**5**	**3**	**5**	4.3	1.2	3	4	4	3.7	0.6	5	5	5	5.0	0.0	4.3	1.0

Table 4: Ratings awarded by subjects 1-6 on 3 occasions of viewing video clips at different levels of foveation

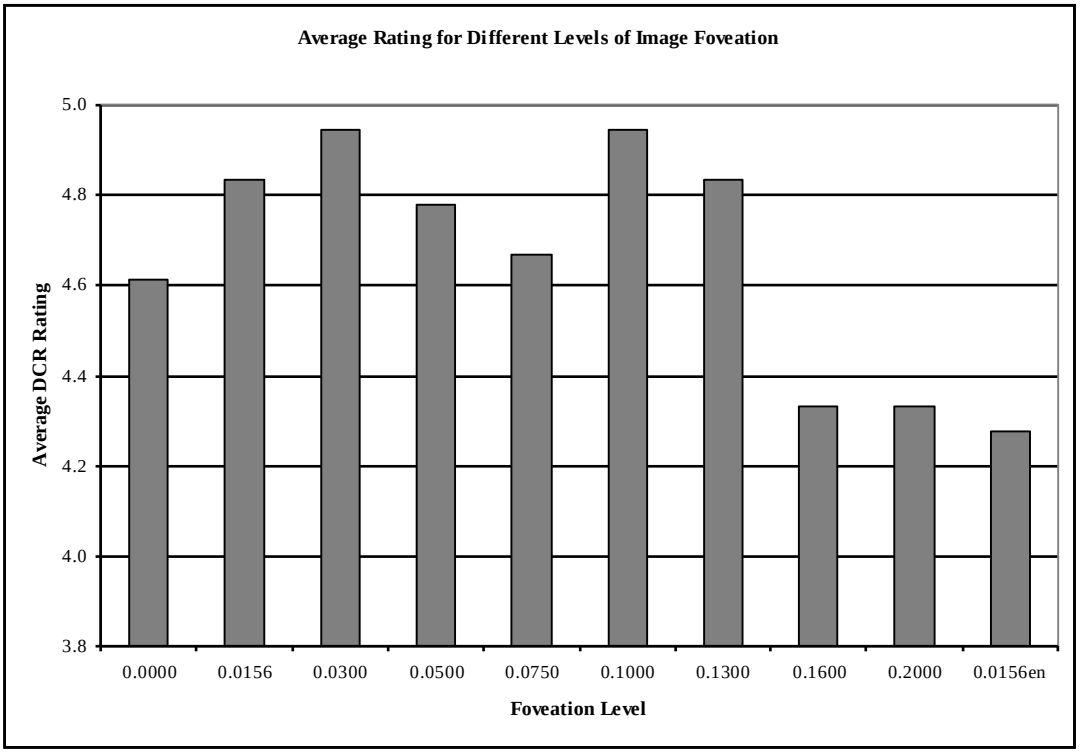

Figure 1: Average MOS for each level of foveation

4 Discussion

The results show that viewer satisfaction is rated high (score 5 or 4) for all conditions tested. A slight drop in the ratings occurs at a foveation levels greater than 0.13. Even at greater degrees of foveation, including the pronounced blurring effect of enhanced foveation (clip 10), only around five to eleven percent of subjects rated the degraded clips as annoying for the task (score 2 or 1). Those subjects (subjects 1 and 2) who rated the degraded clips as affecting sign language understanding (score 2 or 3, none awarded a score of 1) also rated the same clips at a higher score at different points in the experiment. Generally, the deaf subjects appeared to have a very high tolerance level for degraded picture quality as long as the sign language could be understood clearly. Feedback from subject 4 described the degraded picture quality in two of the tests (at 0.03 and 0.075 foveation levels) as being better than the original for sign language comprehension. After the experiment, Subject 1 reported that he had the impression that the signer had shifted position compared to the source clips at foveation levels over 0.075 (he did not mention the significant blurring that occurred at this level of foveation).

Previous research on the use of subjective assessment [12, 17] has questioned the use of ITU recommended scales for subjective assessment of video. The ITU scale is primarily concerned with determining whether subjects can detect degradation in picture quality. They argue that quality evaluation should be related to user task. The problems of subjects accepting low picture quality, particularly if there is an associated notion of cost, and the limitations of the scale are discussed in their research. In our subjective testing experiment the users compared the foveated clip with the original source and so no cost comparisons were implied between clips with different levels of foveation. Our adaptation of the rating scale ensured that the criteria for awarding the rating scores were specifically related to the task of comprehending sign language, rather than a general evaluation of overall picture quality by the viewer. Other approaches, discussed by Wilson and Sasse [17], provide interesting additional subject feedback (for example physiological measurements such as Heart Rate, Blood Volume Pulse and Galvanic Skin Resistance). However, physiological measures of user cost/stress may interfere with the subject's primary task and might also be difficult to separate from other emotions arising from the content of the sign language material.

Our results are encouraging in the sense that it appears that deaf people watch sign language video in the way that Siple described [16] and in the same characteristic manner observed in our previous eye movement tracking experiments (described in the Introduction section of this paper). As long as the face of the signer was displayed in high spatial resolution, the deaf viewers were able to understand the video content, displayed at twenty-five frames per second, even when the peripheral area was significantly blurred.

This leads to the conclusion that there is potential to exploit the viewing behaviour of deaf people in the design or adaptation of video communication systems. Selective prioritisation of important regions of video images may enable more efficient transmission and improve the perceived quality of sign language video content by deaf people. Video coding standards achieve compression using motion compensated prediction followed by transform coding, quantisation and entropy coding [13]. The coding process results in some loss of quality in the decoded video sequence. Increasing the quantiser step size increases compression and reduces decoded video quality. Prioritised coding of sign language video could be achieved by (for example) reducing the quantiser step size in the face region of the image and increasing the step size further away from the face, resulting in higher compression of the regions that are perceived in peripheral vision.

Work is currently in progress to optimise the performance of video communication for deaf people based on visual response mechanisms to sign language video. The aim of this work is to improve perceived video quality at low bit rates (less than 200kbps) and to provide good full-screen DVD quality video on standard systems (256kbps) which currently give 'good quality' quarter-screen (CIF) images. The resulting content-prioritised coding scheme will be tested using a suitable method of subjective testing developed for the task.

Acknowledgements

The authors would like to acknowledge the help and support of Jim Hunter who acted as BSL interpreter and Lisa Davidson who provided the sign language content for the experiment. Special thanks to Edith Ewen and the deaf people at the Aberdeen Deaf Social and Sports Club for their continued interest and support and for taking part in the experiments.

References

[1] [1] D. Agrafiotis, N. Canagarajah, D. R. Bull, J. Kyle, H. Seers and M. Dye, "A video coding system for sign language communication at low bit rates", *Proc. IEEE International Conference on Image Processing*, Singapore, (2004).

[2] 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* (**3**), (1995).

[3] W. S. Geisler, and J. S. Perry, "A Real-Time Foveated Multi-Resolution System for Low Bandwidth Video Communication", *SPIE Proceedings **3299,*** (1998).

[4] ISO/IEC 14496-10 and ITU-T Rec. H.264, "Advanced Video Coding", Geneva: ITU-T (2003).

[5] ITU-T Rec. H.263, "Video Coding for Low Bit Rate Communication", Geneva: ITU-T (1998).

[6] ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications", Geneva: ITU-T (1999).

[7] ITU-T SG16, "Draft Application Profile: Sign Language and Lip Reading Real Time Conversation Usage of Low Bit Rate Video Communication", Geneva: ITU-T, (1998).

[8] S. Kleinfelder, "Foveated Imaging on a Smart Focal Plane", March 19 1999. http://ise.stanford.edu/class/psych221/projects/99/stuartk/fovis.html [accessed 15/01/04]

[9] L. J. Muir and I. E. G. Richardson, "Video Telephony for the Deaf: Analysis and Development of an Optimised Video Compression Product", *Proc. ACM Multimedia Conference*, December, Juan Les Pins (2002).

[10] L. J. Muir, I. E. G. Richardson and S. Leaper 2003 "Gaze Tracking and its Application to Video Coding", *Proc. Int. Picture Coding Symposium*, April, Saint-Malo, (2003).

[11] L. J. Muir and I. E. G. Richardson, "Perception of Sign Language and its Application to Visual Communications for Deaf People" *Journal of Deaf Studies and Deaf Education* (submitted October 2004).

[12] J. Mullin, L. Smallwood et al., "New Techniques for assessing audio and video quality in real-time interactive communications", IHM-HCI, Lille, France, (2001).

[13] I. E. G. Richardson, "H.264 and MPEG-4 Video Compression", Chichester: John Wiley & Sons, (2003).

[14] D. M. Saxe and R. A. Foulds, "Robust Region of Interest Coding for Improved Sign Language Telecommunication", *IEEE Transactions on Information Technology in Biomedicine, **6** (4),* (2002).

[15] R. Schumeyer, E. Heredia and K. Barner, "Region of Interest Priority Coding for Sign Language Videoconferencing", *Proc. IEEE Workshop on Multimedia Signal Processing*, June, Princeton. (1997).

[16] P. Siple, "Visual Constraints for Sign Language Communication", *Sign Language Studies pp* 95-110, (1978).

[17] G. M. Wilson and M. A. Sasse, "Do Users Always Know What's Good For Them? Utilising Physiological Responses to Assess Media Quality", *Proceedings of HCI 2000: People and Computers XIV - Usability or Else!* Sunderland, UK, Springer, (2001).

[18] W. W. Woelders, H.W. Frowein, J. Nielsen, P. Questa and G. Sandini, "New Developments in Low-Bit Rate Videotelephony for People who are Deaf", *J. Speech, Language and Hearing Research*, **40**, pp. 1425-1433, (1997).