Perception of Sign Language and its Application to Visual Communications for Deaf People

Laura J. Muir and Iain E. G. Richardson

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

Abstract

Video communication systems for deaf people are limited in terms of quality and performance, particularly at low bit rates. Analysis of visual attention mechanisms for sign language may enable optimisation of video coding systems for deaf users. Eye movement tracking experiments were conducted with profoundly deaf volunteers while watching sign language video clips designed to test the full range of movements and gestures which make up British Sign Language (BSL). Deaf people are found to look mostly at the face region, with some subjects exhibiting occasional short excursions away from the face. Factors that drive the gaze away from the face are found to be hand gestures near the face and expansive movements of the hands and body in the lower region of the image. The implications of these results for the design of optimised visual communication systems for deaf users are discussed.

Perception of Sign Language and its Application to Video Communications for Deaf People

1.0 Introduction

Visual perception is the process of acquiring knowledge about environmental objects and events by extracting information from the light they emit or reflect [Palmer 2002]. How we 'see' remains an active research challenge for vision scientists and specialists. Understanding the detection, recognition and interpretation of visual information could have a tremendous impact on how we present and use visual information and on the design of information systems. The challenge is to understand how visual information can be presented so that its use can be optimised for the observer.

Of all the senses, vision is relied on most heavily for sensory input about the environment [Hendee and Wells 1997]. This is particularly true for deaf people who rely on visual communication of information using sign language and/or lip reading. The aim of this research is to investigate how deaf people 'see' sign language. The rationale for this is that an understanding of how deaf people observe sign language could enable video communication systems, for example video conferencing, to be optimised.

In this study we examine the influence of sign language video content on the attention mechanisms of deaf viewers and the implications for design of video communications systems for deaf people. We review the quality requirements for sign language video communication and what is known about scene perception and the eye gaze of deaf people observing sign language. An experiment is presented,

using eye tracking, to investigate how deaf people perceive sign language video and discuss this in the context of improving sign language video communication quality.

1.1 British Sign Language and Video Quality Requirements

Sign language is a complex combination of facial expressions, mouth/lip shapes, hand and body movements and finger spelling. Visual communication of information between deaf people during freely expressed sign language conversation is detailed and rapid. Movements of the hands during a period of finger spelling can be observed to be blurred even when captured at 25 frames per second. An ITU-T draft profile [ITU-T SG16 1998] details the quality requirements for sign language video communication including a minimum of CIF resolution (352x288 displayed pixels) and frame rate of at least 25 frames per second. Visual perception of sign language video requires sufficient spatial and temporal resolution to capture the detailed movements of the signer. Reasonable visual quality and frame rates can be obtained using current video compression coding Standards such as H.263 [ITU-T H.263 1998] at high bit rates. At bit rates below c. 200 kilobits per second (kbps), real-time video communication is characterised by low frame rates, small picture sizes and/or poor picture quality [Richardson 2003]. Deaf people using video phones have to make modifications to try to overcome these problems, for example using slow exaggerated movements. This can prove to be tiring and frustrating to the user and limits the usefulness of video technology to the deaf community. Even the improved video compression efficiency of the new H.264 coding standard [ITU-T H.264 2003] may not be acceptable for accurate sign language communication at low bit rates.

Deaf people are enthusiastic about the use of technology for personal communication at a distance but frustrated by the current poor performance at low bit rates characterised by poor picture quality and jerky movements [McCaul 1997]. There is therefore a requirement to optimise video communication systems for deaf users and this motivates the study of perceptual behaviour of deaf people described in this paper.

1.2 Visual Perception of Sign Language

This section reviews what is known about human visual processing and the perception of sign language by deaf people.

It is generally accepted that the examination of a visual stimulus involves parallel pre-attentive processing (first glimpse to give a global impression of the stimulus) and focal attention [Palmer 2002]. Focal attention involves serial scanning of an image using eye movements. Information is processed in detail from the foveal area (in particular, approximately 2.5 degrees of visual angle around the centre of the visual field) and in reduced detail from the larger peripheral area around the fovea. Movements of the head and eye direct the foveal region of high visual acuity to visually sample selected areas of the stimulus.

A saccade is a rapid eye movement which is used to visually scan the scene and bring different areas to fall on the fovea [Cumming 1978]. A saccade requires approximately 150-200ms for planning and execution and reaches an angular velocity of up to 900 degrees/sec. Fixations occur between saccades, during which the eye dwells on an object for a variable period of time. The average duration of a fixation is 300ms [Palmer 2002].

Perceiving a realistic visual scene generally requires a sequence of many different fixations [Findlay and Gilchrist 2003]. Foveal information is clear and fully chromatic whereas peripheral information is blurry and weak in colour to a degree depending on the distance from the fovea. In order to obtain high resolution information about the spatial and or chromatic attributes, the visual scene must be explored using eye movements to place different information in the fovea at different times.

The process of saccadic exploration of complex images was investigated using crude equipment by Yarbus [Yarbus 1967]. He recorded fixations and saccades observed while viewing objects and scenes. By superimposing eye movements on the stimulus picture, he was able to determine which parts of the image observers found most informative. He observed that the way in which the eyes explored a complex image depended on the task. More sophisticated eye movement tracking equipment has allowed researchers to determine the specific sequence of fixations that observers execute when exploring a visual stimulus/scene.

Voluntary eye movements are the main instruments of selective attention. Attention is global (to the whole scene), to a selected object or set of objects, to a specific part of an object or to the property of an object (e.g. colour). Mack and Rock [Mack and Rock 1998] proposed that attention is required for conscious perception of anything at all. Accurate measurement of where an observer is looking is not always a measure of attention [Shepherd et al. 1986]. The human response to a visual stimulus depends on many factors but is ultimately task-specific [Findlay and Gilchrist 2003; Gale 1997], that is, how we see depends on the task being performed.

Land [Land et al, 1999] used eye tracking to investigate eye movements during active tasks including; driving, table tennis, piano playing and tea-making. The results demonstrated that gaze is directed to the points of the scene where most information can be extracted and that the eye anticipates movement rather than follows it. The cognitive aspect of the task was demonstrated to have an important effect on viewing behaviour.

The task of sending and receiving sign language signals was explored by Patricia Siple [Siple, 1978]. Siple proposed that, since sign language is received and initially processed by the visual system then we would expect that the rules for forming signs would be constrained by the limits of that system. She observed that subjects viewing sign language look at the face, with small excursions around the face, of the signer. This behaviour demonstrates the importance of the face in giving cues to the meaning of gestures. Her paper studied the development of the sign system to maximise the information that the eye can gather. In sign language production, small detailed motions were observed to occur in and around the face and upper body region where the receiver (looking at the signers face) can observe gestures in high acuity. Large, less detailed gestures are produced in the peripheral region of view and therefore observed by the receiver at low visual acuity. These large motions tend to be in the vertical and horizontal axes where acuity is greater than for other orientations by the periphery. Siple also described the use of redundancy to further maximise the information that can be conveyed in the peripheral region of view. The conclusion of the study by Siple is that efficient communication of sign language between deaf people has developed within the constraints of the Human Visual System.

The nature of a task is key to viewing behaviour. Eye movements studied in parallel with an articulated theory of cognitive activity for the task in question can provide useful information about visual perception [Vivianni 1990]. Our research investigates the eye movements of deaf people receiving sign language and proposes how video communication systems may be optimised to take account of what is known about the production and observation of signs within human visual limits. We postulate that a deaf person observing sign language is carrying out a specific task that produces a characteristic and consistent pattern of visual attention response which can be exploited to optimise video communication systems such as video telephony and video conferencing systems.

1.3. Video Communication of Sign Language - Previous Research

Previous work on video communication of sign language has been limited by not addressing temporal and spatial quality requirements, visual perception mechanisms and by a lack of consultation and testing with the target user group (i.e. deaf people).

The effect of frame rate and spatial resolution on speech reading (mouth/lip shapes), finger spelling and gestures was investigated by Woelders et al. [Woelders et al. 1997] who demonstrated that frame rate had a significant effect on the communication of mouth shapes in particular. The resulting compressed video from this and other research in this field [Schumeyer et al. 1997; Eleftheriadis and Jacquin 1995] produced distorted images which were not subject to quality testing by the target end user. Image segmentation and region of interest coding schemes have been proposed, for example skin detection [Saxe and Foulds 2002]. Other methods include

foveated processing to mimic human visual processing [Geisler and Perry 1998]. However, these methods produce distorted images and are based on the assumption that the hands must be transmitted at the same (temporal/spatial) quality as the face. None of the research available in the literature prior to 2002 considers the perceptual responses of deaf people watching sign language.

Our initial gaze tracking experiments with eight deaf volunteers [Muir and Richardson 2002; Muir et al. 2003] established that sign language users exhibit a consistent characteristic eye movement response to sign language video. The results of these experiments (confirmed independently by [Agrafiotis et al. 2003]) support the theory by Siple [Siple 1978] that deaf people perceive the face in high visual resolution and that hand gestures are viewed in peripheral, lower resolution, vision.

1.4 Experimental Design and Rationale

The investigation presented in this paper builds on our previous work [Muir et al. 2003] and extends the study to include more participants, a wider range of sign language video material and more detailed analysis of the factors that influence the attention of deaf people watching sign language video.

Eye tracking was used to explore the visual response of deaf participants to video stories which were selected to include a wide range of fine and gross sign language movements and gestures. Eye tracking is a fast and accurate method of capturing and processing gaze data; in this experiment, temporal resolution up to 60Hz and internal processing up to 640x480. It permits investigation of on-line processing of full-screen video images and does not disrupt normal viewing.

In the experiment, profoundly deaf adult participants, who used BSL as their first language, observed three short video clips with the task of understanding the signed stories in each clip. Eye movement data was captured and compared for each subject and clip.

Experiments were conducted under controlled conditions in a room with 100% artificial, overhead lighting. The subject was positioned at a comfortable viewing distance (4 to 6 times the screen height) from the monitor.

Each subject was given instruction in British Sign Language (BSL) through a qualified interpreter. All communication with the subject was in BSL, the subjects' first language, no printed instructions or feedback forms in English language were used.

The results of the eye movement tracking experiments were analysed, for each participant, by playing back the video clip and plotting the recorded (x, y) eye position co-ordinates on each video frame. The gaze points were examined frame-by-frame with respect to the designated areas of the video image. The selected areas were; upper and lower face, hands, fingers, upper body, lower body, background and object (a camera on a tripod in clip one). These were chosen so that the researcher could identify the most important regions of the scene for sign language communication. The distinction between upper and lower face was made to determine if the region around the eyes (upper face) or around the mouth (lower face) was more significant for sign language understanding. The distinction between hands and fingers was made to test whether wide movements of the hands and detailed movement of the fingers (for example, during finger spelling) were followed by the viewer. The upper body area was defined as the area below the chin and above the

waist of the signer and the lower body was defined as the area below the waist. A fixation was recorded as a gaze of duration of 0.03 seconds or more [Palmer 2002]. In cases where the regions over-lapped (for example, when the hands where over the face region), the sequence of eye movements before and after this occurrence was observed to estimate which region was being followed by the eye. The data was analysed in two ways. Firstly, the total fixation time on each of the designated regions was recorded to determine which region was most important to the viewer. Each subject's fixation time was expressed as a percentage of the total viewing time for each clip to allow comparison between viewers and to compare the results for each video clip. Secondly, a timeline was produced, for each subject, which recorded the location of fixations, during each of the videos, with respect to time. The data was examined on a frame-by-frame basis and the gaze point noted with respect to the sign language action in the video. Figure 1 includes an extract from the timeline for video clip 2. It shows the gaze locations of each of the ten subjects for the first five seconds of the video clip. The gestures are noted along the top row of the timeline and colour-coded to match the colours used to represent the designated regions of the image. Sample frames from the video clip are included in the Figure to illustrate the video content.

2.0 Method

2.1 Subjects

Eye movement tracking experiments were conducted with seventeen profoundly deaf-from-birth volunteers from the Aberdeen Deaf Social and Sports Club (ADSSC). For each subject British Sign Language (BSL) was their first language and English was a second language. For this reason all communications were in BSL, aided by a local British Sign Language interpreter who was known to the participants.

Seven of the participants were excluded from the experiment as we were unable to obtain consistent accurate tracking of their eye movements during calibration. Of the ten subjects proceeding to the experiment, seven were male and three were female and ages ranged from thirty to eighty-two years.

2.2 Apparatus

Eye movements were captured by a ViewPoint eye tracker from Arrington Research Ltd. (Cambridge, U.K) incorporating an infra-red light source and camera mounted on a clamp with a nose bridge and chin rest for comfortable and secure positioning of the subject's head. The infra-red light source illuminates the eye and provides reflection from the smooth cornea. The camera captures the video signal, reflected light from the eye, which is digitised by a video capture device in the PC. Image segmentation algorithms are applied to the digitised image to locate the dark pupil of the eye. Eye position signals are transformed to produce eye movement co-ordinates. Data gathered from a calibration routine, before the test begins, is used to calculate the point of regard.

Video clips were displayed to the viewer on a seventeen inch monitor (A) with true colour, 32 bit display connected to a Dell Pentium IV PC with PCI Video Capture Card installed. A second monitor (B) was connected to the PC (not visible to the subject) for the researcher to control and monitor the experiment.

2.3 Materials

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 two profoundly deaf volunteers. The volunteers were from the same geographical area, the North-East of Scotland, and used the same version of BSL as the subjects participating in the experiment. It is worth noting that BSL has regional variations analogous to speech dialects. The signers in the video related short stories from their own experience using their own natural style and expression of signing. Three video clips were selected to ensure the test material contained a wide range of sign language movements, expressions and gestures (including finger spelling) as described below.

The first clip (22.08 seconds) displays a close view of the signer (from the waist upwards). The signer used facial expression, lip movement and gestures but limited body movement around the scene which also included a camera and ventilation shaft. These background objects were included to test whether they would prove to be a distraction for the viewer. The story told in this clip is of the signer's experience of communication between deaf and hearing members of her family. An English translation of her story is; "A long time ago, when I was young, I would ask my mother what everyone was saying. Now when my children speak with no voice,

my mother asks me what they are saying. I remind her that she wouldn't tell me what was being said until they were finished and so she will just have to wait too. She realises this now".

In the second clip (27.20 seconds) the signer (same as in clip one) is at a greater distance from the camera and seen above knee height. The signer used facial expression, lip movement, big gestures and detailed finger-spelling but limited body movement around the scene which had no distracting objects. The story told in this clip is of the signer's experience as a child at school learning to use her voice. An English translation of her story is; "When I was at school, a long time ago, when I was small, my speech was hopeless. They tried to teach me to speak but it just went over my head. The teacher said it was a bit of a problem. She said some people are good but you are not good, your speaking is not good. So I had to lie down on the floor and say 'Ah'. She put a darning needle in my mouth to make the 'A'sound. My heart was throbbing."

In the third clip (46.64 seconds) the signer used facial expression, lip movement, finger-spelling, wide gestures and movement around the scene to tell the story of his experience on holiday. An English translation of his story is; "We have been to America, three times, and also to Spain. We met deaf people in America. The sign language was different but we could catch certain things by gesturing and so on. Things like 'walking', 'hot', 'drinking' and 'good' we could communicate - and also by writing things down. When I was a boy, I played football and so I could make conversation about that. The language was different - it was interesting."

2.4 Procedure

The eye tracker camera was set up so that the video image of the subject's pupil (dominant eye where appropriate) was in the centre of the control display window in monitor B. The tracking system was adjusted in set-up mode (temporal resolution = 30Hz, internal processing = 340x240) so that the threshold area of the dark pupil of the eye and the white corneal reflection was obtained in the search area. The scan density was adjusted to obtain the minimum number of points which would correctly locate the dark pupil for maximum possible accuracy. Following the set up stage, the equipment was calibrated for the individual subject to obtain co-efficients for internal mathematical mapping. Calibration was performed at temporal resolution = 30Hz, internal processing = 640x480 to obtain the highest possible degree of accuracy. The subject was instructed to foveate on each of sixteen calibration points on monitor A, until they disappeared from the screen, avoiding anticipation of the next point. The researcher controlled and monitored the calibration routine on monitor B, checking the success of calibration and re-presenting the stimuli as required. Once calibrated, the video stimuli (three videos separated by further calibration markers) were presented full-screen to the subject on monitor A. Eye movements were processed at temporal resolution = 60Hz, internal processing = 340x240 and monitored by the researcher on monitor B. The (x,y) co-ordinates of the captured gaze data were saved to a unique data file for each subject.

The total time for the experiment with an individual participant was approximately twenty minutes. At the end of the experiment, subjects were asked if there was anything in the sign language video that could not be understood, was not clear or needed to be repeated. The rationale for an open ended question unrelated to

the video content was that the researchers wished to test ease of relaxed, natural sign language communication to the subject rather than test comprehension which might have influenced the way the video clips were regarded.

3.0 Results

All subjects reported ease of sign language communication with no requests for clarification or repetition. Conversations with the subjects after the experiment, through the BSL interpreter, demonstrated understanding of, and interest in, the content of the video clips used.

3.1. Fixation on Regions of Importance

The total fixation time (seconds) in separate designated regions of the video image was recorded for each of the 10 subjects. Total fixation times for each subject vary depending on the number of saccades during viewing. The total and percentage fixation times for each subject, for each of the test video sequences, are given for each region of the video image in Tables 1 – 3. The tables also show the average total and percentage time spent looking at each of the designated image regions.

Table 1: *Total and Percentage Fixation Times on Different Regions of Video Clip 1*

	Upper Face		Lower face		Hands		Fingers		Upper Body		Lower Body		Object		Background	
Subject	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%
1	**21.44**	**97.99**	**0.28**	**1.28**	**0.00**	**0.00**	**0.00**	**0.00**	**0.04**	**0.18**	**0.12**	**0.55**	**0.00**	**0.00**	**0.00**	**0.00**
2	7.48	33.98	5.92	26.90	1.61	7.31	0.00	0.00	6.68	30.35	0.00	0.00	0.32	1.45	0.00	0.00
3	4.92	22.99	4.60	21.50	0.80	3.74	0.00	0.00	11.08	51.78	0.00	0.00	0.00	0.00	0.00	0.00
4	**21.56**	**98.54**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.32**	**1.46**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
5	**21.52**	**98.35**	**0.36**	**1.65**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
6	16.16	78.91	4.32	21.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	**21.08**	**96.17**	**0.84**	**3.83**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
8	17.52	87.08	2.08	10.34	0.08	0.40	0.00	0.00	0.08	0.40	0.00	0.00	0.36	1.79	0.00	0.00
9	**18.78**	**90.03**	**1.64**	**7.86**	**0.44**	**2.11**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
10	3.12	18.16	11.42	66.47	2.04	11.87	0.00	0.00	0.60	3.49	0.00	0.00	0.00	0.00	0.00	0.00
Average	15.36	72.22	3.15	16.09	0.50	2.54	0.00	0.00	1.88	8.77	0.01	0.05	0.07	0.32	0.00	0.00
St.Dev.	6.96	31.63	3.38	19.10	0.72	3.85	0.00	0.00	3.64	16.88	0.04	0.16	0.14	0.65	0.00	0.00
Average	**20.88**	**96.22**	**0.62**	**2.92**	**0.09**	**0.42**	**0.00**	**0.00**	**0.07**	**0.33**	**0.02**	**0.11**	**0.00**	**0.00**	**0.00**	**0.00**
St.Dev.	**1.06**	**3.21**	**0.58**	**2.76**	**0.18**	**0.84**	**0.00**	**0.00**	**0.12**	**0.57**	**0.05**	**0.22**	**0.00**	**0.00**	**0.00**	**0.00**

Table 2: *Total and Percentage Fixation Times on Different Regions of Video Clip 2*

	Upper Face		Low er face		Hands		Fingers		Upper Body		Low er Body		Background	
Subject	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%
1	6.44	24.66	15.88	60.80	0.84	3.22	0.24	0.92	2.60	9.95	0.12	0.46	0.00	0.00
2	5.88	24.30	3.52	14.55	2.36	9.75	0.32	1.32	12.12	50.08	0.00	0.00	0.00	0.00
3	7.24	27.93	10.20	39.35	2.16	8.33	0.00	0.00	6.32	24.38	0.00	0.00	0.00	0.00
4	8.00	30.40	14.16	53.80	1.24	4.71	0.00	0.00	2.92	11.09	0.00	0.00	0.00	0.00
5	23.16	85.91	3.52	13.06	0.00	0.00	0.00	0.00	0.28	1.04	0.00	0.00	0.00	0.00
6	22.00	87.58	2.88	11.46	0.00	0.00	0.00	0.00	0.24	0.96	0.00	0.00	0.00	0.00
7	26.20	97.76	0.28	1.04	0.00	0.00	0.00	0.00	0.32	1.19	0.00	0.00	0.00	0.00
8	9.48	38.23	13.00	52.42	0.08	0.32	0.92	3.71	1.16	4.68	0.16	0.65	0.00	0.00
9	3.00	11.45	14.56	55.57	5.28	20.15	0.00	0.00	3.36	12.82	0.00	0.00	0.00	0.00
10	9.72	37.44	13.72	52.85	0.00	0.00	2.20	8.47	0.32	1.23	0.00	0.00	0.00	0.00
Average	12.11	46.56	9.17	35.49	1.20	4.65	0.37	1.44	2.96	11.74	0.03	0.11	0.00	0.00
St.Dev.	7.91	29.70	5.64	21.65	1.61	6.22	0.67	2.60	3.57	14.62	0.06	0.22	0.00	0.00

Table 3: *Total and Percentage Fixation Times on Different Regions of Video Clip 3*

	Upper Face		Lower face		Hands		Fingers		Upper Body		Lower Body		Background	
Subject	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%
1	39.04	86.83	2.84	6.32	0.76	1.69	0.00	0.00	1.80	4.00	0.52	1.16	0.00	0.00
2	3.24	8.09	11.68	29.18	4.19	10.47	0.00	0.00	20.92	52.26	0.00	0.00	0.00	0.00
3	27.72	60.84	2.96	6.50	0.00	0.00	0.00	0.00	13.04	28.62	1.84	4.04	0.00	0.00
4	4.16	9.04	2.68	5.82	0.00	0.00	0.00	0.00	39.20	85.14	0.00	0.00	0.00	0.00
5	29.36	64.61	14.72	32.39	0.00	0.00	0.00	0.00	1.36	2.99	0.00	0.00	0.00	0.00
6	17.24	40.87	21.26	50.40	0.00	0.00	0.00	0.00	3.68	8.72	0.00	0.00	0.00	0.00
8	20.72	52.24	4.76	12.00	3.04	7.67	0.00	0.00	11.14	28.09	0.00	0.00	0.00	0.00
9	21.72	48.40	10.32	22.99	0.80	1.78	0.00	0.00	12.04	26.83	0.00	0.00	0.00	0.00
10	2.24	4.84	0.96	2.07	0.00	0.00	0.00	0.00	43.12	93.09	0.00	0.00	0.00	0.00
Average	18.38	41.75	8.02	18.63	0.98	2.40	0.00	0.00	16.26	36.64	0.26	0.58	0.00	0.00
St. Dev.	12.22	27.16	6.50	15.31	1.47	3.69	0.00	0.00	14.58	31.61	0.58	1.28	0.00	0.00

The average percentage fixation times are plotted in Figure 2 to allow comparison of the results obtained for the three video clips used in the experiment.

The results for video clip one (Table 1) demonstrate that, on average, most of the time was spent looking at the face (88.31%) and in particular the upper face region (72.22%) of the video. Subjects 1,4,5,7,and 9 (shown in bold type in Table 1) displayed a very similar pattern of viewing times and looked almost exclusively at the upper face during this video clip (96.22% with low standard deviation). Subjects 6 and 8 exhibited behaviour similar to this group in terms of the time spent looking at the face although their gaze fell on the lower face more than the rest of the group (21.09% and 10.34% respectively). The average fixation time on lower body of the signer and the background object (camera) was less than the threshold time for a fixation. Four of the subjects (subjects 2,3,9 and 10) spent an average of 0.5 seconds (2.54%) of the total viewing time looking at the hands and four of the subjects (subjects 2,3,4 and 10) spent an average of 1.88 seconds (8.77%) of the total viewing time looking at the lower body region.

The results for video clip two (Table 2) show that most of the fixation time (82.05%) was on the face region. Subjects 5, 6, and 7 exhibited similar behaviour to that for clip 1 (average of 90.42% of fixation time on the upper face). Subjects 1, 4, 8,9,10 spent more time (an average of 55.09% fixation time) looking at the lower face region. Subjects 2 and 3 show a similar viewing pattern to that shown for clip 1, that is fixating more on the upper body region.

Results for video clip three are shown in Table 3 (data for subject 7 is excluded as he was the signer in the video clip). The average time spent looking at the face in this test was 60.38% of the fixation time. More of the fixation time,

36.64% on average, was spent looking at the upper body region which includes the area just below the face and the chest of the signer. Three of the subjects (subjects 2, 4 and 10) spent most of their fixation time on the upper body (in contrast to the behaviour of subjects 4 and 10 during sequences 1 and 2).

Plotting the average fixation times on the designated areas, for each of the clips, (Figure 2) shows similar curves (patterns of viewing behaviour) but in varying proportions. The difference in the pattern of results obtained for the three clips is explored further to determine if the data could have come from the same population (null hypothesis) or if the difference between at least two of the data sets is statistically significant.

3.1.1. Statistical Comparison of Viewing Behaviour for Three Video Clips

A non-parametric, Friedman Test was conducted to determine whether there was a statistically significant difference in the percentage fixation times on the specified regions of each of the test video sequences by the subjects in the sample at the 5% significance level. A non-parametric test is applied to ordinal or interval data, is distribution free and tests whether population locations differ [Keller and Warrack 2003]. The eye location data is interval (percentage fixation times) and is not normally distributed. The null hypothesis for the test is that the data for all 3 clips could have come from the same population and are not significantly different.
The Friedman test ranks the results (percentage fixation times) for the subjects for each video and uses chi-squared distributions to determine whether at least two of the data sets differ. The SPSS output is as follows:

Ranks

	Mean Rank
Video 1	2.56
Video 2	1.89
Video 3	1.56

Test Statistics[a]

N	9
Chi-Square	4.667
df	2
Asymp. Sig.	.097

a. Friedman Test

The test significance result is a 0.097 probability that there is no significant difference in the results obtained for the three videos. This is greater than the level of significance (0.05 probability). The Friedman test indicates that, for this sample, there was no statistically significant difference in the viewing behaviour of subjects for the three different types of video sequence used in the experiments.

3.2 Fixation in Relation to Video Content

Further examination of the raw data was conducted to explore the motivating factors for eye movements. The sequence of fixations for each subject was examined with respect to the sign language content of each video clip. A timeline (similar to the example shown in Figure 1) was produced for each subject.

In the first video clip, the short excursions to the hands exhibited by subjects 2, 3, 8, 9 &10 were found to be associated with movement of the hands near to (to one side of) the face region in the sign language video, possibly because the hands were close enough to “draw” the eyes away from the face but still allow the face to be seen at high resolution. Two of the subjects (subjects 2, 3) spent a greater percentage of

their total fixation time looking at the upper body region (30.35% and 51.78% respectively) which included the area just below the face. Examination of the timeline suggested that the gaze of these subjects was closer to the location of the hands than the other participants.

Motivating factors taking gaze away from the face in video clip 2 were investigated by examining the timelines for each subject. Gaze away from the face (mostly to the upper body region) occurred during pauses in sign language and when gestures and movements were located in the lower body region of the signer. None of the subjects followed the hands or fingers during the periods of finger spelling in the video. Gaze was found to be in the (upper or lower) face region during finger spelling in all cases.

Examination of the timelines for video clip 3, indicated that factors influencing gaze in the upper body region were large gestures (in the lower body region of the signer) and movement of the signer around the scene, particularly towards the end of the clip.

The results imply that the face is the centre of attention for a deaf person observing sign language, particularly for sequences where the signer uses a range of gestures and finger spelling but without wide ranging body movements (video clips 1 and 2). Gaze is mostly in the upper face region for clip 1 (in which there is a closer view of the signer) and more time is spent on the lower face in clip 2 (in which the signer is further away from the camera, makes wider gestures and uses more detailed finger spelling). Hand gestures close to the face, expansive gestures in the lower body region of the signer and movement of the signer around the video scene were found to

act as “drivers” (motivating factors), taking the subject’s gaze away from the face region.

Discussion

The aim of this investigation was to explore how profoundly deaf people view sign language video content and the application of this to the design of video communication systems.

In the introduction to this paper we identified the importance of the task and the nature of the sign language material on gaze patterns. The work of Siple [Siple 1978] was important for understanding the relationship between the Human Visual System and the development and production of sign language. What is seen in clear foveal vision and the information that can be gathered from peripheral vision can be used to guide the development of systems (sign systems or video systems) that work optimally within the limitations of human vision.

Our eye movement tracking experiment was designed to test the responses of deaf viewers to a wide range of sign language movements and gestures and to investigate viewing patterns that might be exploited in the design of optimised video communication systems. Our results demonstrate that the most important region of the sign language video image is the face of the signer. This is particularly evident in the results obtained for video clip one where the signer is closer to the camera than in the other video clips. Fixations are mainly on the upper face region with no visual excursions to the distracter objects in the background. Gaze is more on the lower face region for video clip two where the signer is further from the camera and the face region is therefore smaller. Participants were found not to follow the movements of

the hands or detailed movements of the fingers during periods of finger spelling, suggesting that sign information was observed in peripheral (lower resolution) vision. Short excursions to the hands were noted only when the hands of the signer were close to the face. The hands were close enough for the face to remain in foveal (high resolution) vision. The wider, more rapid gestures and movements of the signer in video clip three seemed to cause gaze to fall more on the upper body region of the signer for some viewers. There was no statistically significant difference in the patterns of viewing behaviour across the three videos tested, as determined by the Friedman Test (in section 3.1.1) of this paper. This leads us to conclude that the same viewing strategies are applied by viewers to different aspects of sign language video regardless of the background, distance of the signer from the camera and movement of the signer around the scene.

These findings are supported by vision theory and published research, in particular the previously mentioned work of Siple in relation to sign language. Human perception of motion is an important factor which may influence the way deaf people view sign language. It has been demonstrated that temporal properties of vision are similar across the human visual field [Virsu et al 1982]. As discussed earlier in this paper, the same is not true for spatial vision. Foveal vision (corresponding to a visual angle of 2.5 degrees from the point of fixation) is an area of acute vision. It is the most spatially sensitive part of the visual field, providing high resolution vision.

Extra-foveal, or peripheral, low resolution vision has been shown to have an important role to play in the perception of motion. A study of eccentricity-dependence of motion perception by Baker and Braddick [Baker and Braddick 1985]

concluded that peripheral vision is superior for processing visual motion. They studied the ability of subjects to report the direction of apparent motion when an array of random dots was displaced in relation to retinal eccentricity factors. They found that peripheral vision is specialised for motion perception. They also established that the range of velocities that can be processed increases greatly in peripheral vision whereas in central foveal vision only a very restricted range of velocities could stimulate a vision response.

From our results, detailed spatial vision of the face region was found to be important for comprehension of sign language. Assuming that the hands of the signer 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 vision was found to be adequate for the gross and rapid sign language movements of the hands and body that occurred away from the face region of the signer in our experiment.

We conclude from this that a deaf viewer fixates mostly on the facial region of the signer to pick up small detailed movements, associated with facial expression and lip shapes, which are known to convey important sign language information to the receiver. Small movements of the hands in front of or near to the face can be observed in the foveated region of view but more detailed movement near to the face was found to draw the eyes of some subjects away from the face for a short time. During this time the face was still close enough to be seen in high visual acuity. A deaf person uses peripheral vision to process information from larger, rapid movements of the signer. Fixation on the upper body region (including the area

below the face) by some subjects may have occurred to permit a range of smaller movements to be processed at the edge of the foveal area while still keeping the lower part of the face in high-resolution foveal vision.

These results have a number of implications for visual communication systems. A deaf person requires high spatial resolution in the face region of the signer while temporal resolution is maintained across the entire video scene. This indicates that there is scope for prioritised transmission of sign language video, for example by coding different parts of the scene with varying image quality. It may be possible to reduce the quality of the peripheral region, including body and hands (when away from the face), in a coded video sequence while maintaining perceived video quality. For example, popular video coding standards such as MPEG-4 Visual, H.263 and H.264 achieve compression by a process of motion compensated prediction followed by transform coding, quantization and entropy coding [Richardson 2003]. The coding process is 'lossy', that is, there is some loss of quality in the decoded video sequence. A large quantizer step size produces high compression and poor decoded quality and vice versa. Prioritised coding of sign language video could be achieved by reducing the quantizer step size in the face region and increasing the step size further away from the face, resulting in higher compression of the regions that are perceived in peripheral vision. Extending this priority region to just below the face could enable viewers who need to increase their region of detection of small movements, while maintaining detail for oral sign language signals, could be achieved. The region of clarity for small slow movements could be set for the individual user to allow customisation, as it is clear from the results that content is not always viewed in precisely the same way. Video compression, optimised in this way to meet the needs

of the user, would improve perceived video quality at low bit rates, that is, less than 200 kilobits per second (kbps). Standard systems with bit rates of 256kbps currently giving 'good quality' quarter-screen (CIF) video could be optimised to provide good full-screen DVD quality video images.

Further work is being conducted to quantify the relative requirements for image quality in the regions of a coded sign language video sequence. Tests are in progress to determine the effect of selective coding of sign language video content on perception of quality by deaf people. Part of this work includes the development of a suitable method of measuring subjective quality since standardised methods such as [ITU-T P.910 1999] may not be appropriate for the task-specific nature of sign language video.

The findings presented in this paper demonstrate the potential to exploit the viewing behaviour of deaf people in the design or adaptation of video communication systems for this user group. Selective prioritisation of important regions of the video image may enable more efficient transmission and improve the perceived quality of sign language video content by deaf people.

Acknowledgements

The authors would like to acknowledge the help and support of Jim Hunter who acted as BSL interpreter and organised volunteers for the experiments. 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 eye movement tracking experiments.

References

AGRAFIOTIS, D, CANAGARAJAH, N., BULL, D., DYE, M., TWYFORD, H.E., KYLE, J.G. AND CHUNG HOW, J. 2003. Optimized sign language video coding based on eyetracking analysis. *Proc. Visual Communications and Image Processing*, July, University of Italian Switzerland, Lugano, Switzerland.

BAKER, C.L. AND BRADDICK, O.J. 1985. Eccentricity-dependent scaling of the limits of short-range motion perception. *Vision Research: 25, 803-12.*

CUMMING, G.D. 1978. Eye Movements and Visual Perception *E.C. Carterette, M.P.Friedman (eds), Handbook of Perception: 221-255*. MA: Academic Press.

ELEFTHERIADIS, A. AND JACQUIN, A. 1995. Automatic Face Location Detection and Tracking for Model-Assisted Coding of Video Teleconferencing Sequences at Low Bit Rates. *Signal Processing: Image Communication* (3).

FINDLAY, J. M. AND GILCHRIST, I. D. 2003. *Active Vision: The Psychology of Looking and Seeing.* Oxford: Oxford University Press.

GALE, A.G. 1997. Human Response to Visual Stimuli. *W.R. Hendee, P.N.T. Wells (eds), The Perception of Visual Information*. New York: Springer-Verlag.

GEISLER, W.S. AND PERRY, J.S. 1998. A Real-Time Foveated Multi-Resolution System for Low Bandwidth Video Communication. *SPIE Proceedings Vol. 3299*.

HENDEE, W.R. AND WELLS, P.N.T. (eds) 1997. *The Perception of Visual Information (2nd Edition)*. New York: Springer-Verlag.

ISO/IEC 14496-10 AND ITU-T REC. H.264, 2003. *Advanced Video Coding*. Geneva: ITU-T.

ITU-T REC. H.263, 1998. *Video Coding for Low Bit Rate Communication*. Geneva: ITU-T.

ITU-T REC. P.910, 1999. *Subjective Video Quality Assessment Methods for Multimedia Applications*. Geneva: ITU-T.

ITU-T SG16 1998. *Draft Application Profile: Sign Language and Lip Reading Real Time Conversation Usage of Low Bit Rate Video Communication*. Geneva: ITU-T.

KELLER, G. AND WARRACK, B. 2003. *Statistics for Management and Economics*, 591-594. Thomson Learning.

LAND, M. F., MENNIE, N. AND RUSTED, J., 1999. The Roles of Vision and Eye Movements in the Control of Activities of Everyday Living, *Perception*, 28, 1311-28.

MACK A. AND ROCK, I. 1998. *Inattentional Blindness*. MA: MIT Press.

MCCAUL, T. 1997. Video-Based Telecommunications Technology and the Deaf Community. *Report of Australian Communication Exchange*. Queensland: Australian Communication Exchange Ltd.

MUIR, L.J. AND RICHARDSON, I. E. G. 2002. Video Telephony for the Deaf: Analysis and Development of an Optimised Video Compression Product. *Proc. ACM Multimedia Conference*, December, Juan Les Pins.

MUIR, L.J., RICHARDSON, I. E. G. AND LEAPER S. 2003. Gaze Tracking and its Application to Video Coding. *Proc. Int. Picture Coding Symposium*, April, Saint-Malo.

PALMER, S.E. 2002. *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.

RICHARDSON, I. E. G. 2003. *H.264 and MPEG-4 Video Compression*. Chichester: John Wiley & Sons.

SAXE, D.M. AND FOULDS, R.A. 2002. Robust Region of Interest Coding for Improved Sign Language Telecommunication. *IEEE Transactions on Information Technology in Biomedicine, Vol. 6 (4)*.

SCHUMEYER, R., HEREDIA, E. AND BARNER, K. 1997. Region of Interest Priority Coding for Sign Language Videoconferencing. *Proc. IEEE Workshop on Multimedia Signal Processing*, June, Princeton.

SHEPHERD, M., FINDLAY, J.M. AND HOCKEY, R.J. 1986. The Relationship between Eye Movements and Spatial Attention. *Quarterly Journal of Experimental Psychology*, *38A*: 475-491.

SIPLE, P. 1978. Visual Constraints for Sign Language Communication. *Sign Language Studies* 95-110.

VIRSU, V., ROVAMO, J., LAURENEN, P. AND NASENEN, R. 1982. Temporal Contrast Sensitivity and the Cortical Magnfication Factor. *Vision Research: 22, 1211-1217.*

VIVIANNI, P., 1990. Eye Movements in Visual Search. Cognitive, Perceptual and Motor Control Aspects. In *Eye Movements and Their Role in Visual and Cognitive Processes* (Ed. E. Kowler), Elsevier, Amsterdam pp.353-93.

WOELDERS, W.W., FROWEIN, H.W., NIELSEN, J., QUESTA, P. AND SANDINI, G. 1997. New Developments in Low -Bit Rate Videotelephony for People who are Deaf. *J. Speech, Language and Hearing Research, Vol. 40*, 1425-1433.

YARBUS, A. L. 1967. Eye Movements and Vision. *S.E. Palmer, Vision Science: Photons to Phenomenology*. MA: MIT Press.

Figure Captions

Figure 1. Extract from Timeline for Video Clip 2 (with sample frames)

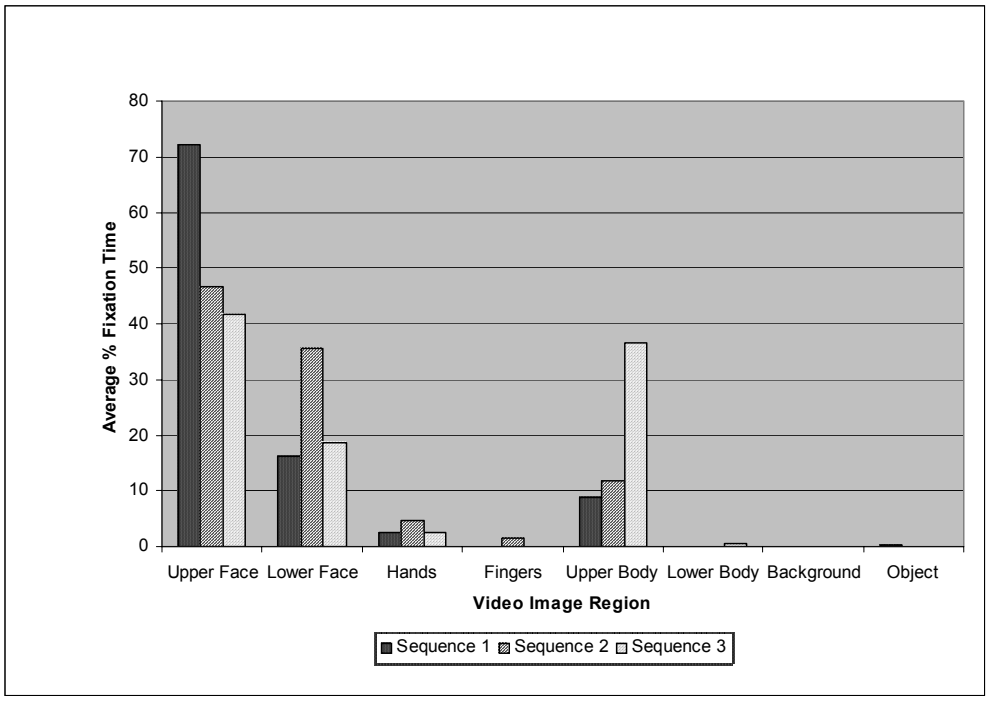
Figure 2. Fixation on Designated Regions of three Video Clips.




	Gesture	Hand gestures at upper body					Hand gestures at lower face	Gestures at UB looking left	R.hand at right UF	Point out from LF to left	R.hand sign at mouth (LF)		Thumbs out at chest		point to eye
	Time (sec)	0		1			2		3		4		5		
Subjects	1	UB	UF				LF				UF		LF	UB	
	2	H			UB		UF		H	LF	UB				
	3	UB	UF							LF	UF				
	4	UB	LF						UF						
	5	UB	UF										LF		
	6	UB	UF				LF	UF	LF		UF				
	7	UB	UF							LF	UB	LF	UF		
	8	UB	UF												
	9	UB	UF												
	10	UB	LF			UF	LF	UF		LF		UB	LF		

KEY:

UF	Upper Face/Eyes
LF	Lower face/Mouth
UB	Upper Body
H	Hands


80
70
60
50
40
30
20
10
0
Average % Fixation Time
Upper Face
Lower Face
Hands
Fingers
Upper Body
Lower Body
Background
Object
Video Image Region
Sequence 1
Sequence 2
Sequence 3