SignWave: Human Perception of Sign Language Video Quality as Constrained by Mobile Phone Technology

Anna Cavender, Erika A. Rice, Katarzyna M. Wilamowska
Computer Science and Engineering
University of Washington
Seattle, WA 98195
{cavender, erice, kasiaw}@cs.washington.edu

ABSTRACT

Many people have benefited from the flexibility that mobile phones provide. The Deaf Community currently has limited access to the mobile telephone network through text messaging systems. However, these force the user to communicate in English as opposed to the preferred language – American Sign Language (ASL). Mobile video phones have the potential to give deaf people the independence and freedom of mobile communication in their indigenous language.

Mobile phone technology offers unique challenges for real time video transfer. Two limitations of this technology are small displays and low network bandwidth. Motivated by these two constraints, we conducted a small study with members of the Deaf Community to determine the video quality needed for effective communication.

We found that general compression techniques do not achieve intelligible ASL under the constraints imposed by current mobile phone networks. More visual movement is required for videos of lower quality, even though objects in the video maintain visual significance regardless of video size or distortion. Screen size was found to be less important to viewer comprehension than distortion.

INTRODUCTION

Cell phones with LCD displays and the ability to transmit and play videos are rapidly becoming more popular and more widely available. Their presence in the marketplace could give deaf and hard of hearing people access to the portable conveniences of the wireless telephone network through the use of sign language.

Figure 1. A hypothetical mobile video phone that people might use to communicate with sign language.

The ability to transmit video (as opposed to text or symbols) would give members of the Deaf Community the most efficient and personal means of remote communication. Some members of the Deaf Community currently use text messaging, but it is extremely cumbersome and impersonal because (a) English is not the native language of most deaf Americans (ASL is their preferred language), and (b) text messaging is much slower than signed conversations. Many deaf Americans use video relay services (where a remote interpreter translates video sign language to spoken English), but this requires equipment (a computer, camera, and internet connection) that is generally set up in the home or work place and does not scale well for mobile use. Video cell phones could potentially make the mobile phone network more universally accessible to over one million deaf or hard of hearing people.

Unfortunately, the Deaf Community in America cannot yet take advantage of this new technology because even today's best video encoders cannot produce the quality video needed for intelligible (ASL) in real time given the bandwidth constraints of the wireless telephone network.

In order for deaf people to utilize video cell phones for mobile communication, a new and different video compression scheme is needed. Given the network constraints, it seems likely that a successful compression scheme will be specific to sign language. Thus, information about the ways in which people view sign language videos will be crucial to the creation of such compression methods.

The purpose of this research is to investigate the effects of screen size and distortion (due to compression) on visual perception of sign language videos. We used an eye tracker to collect eye movement data while members of the Deaf Community watched several videos of varying size and distortion. We also collected subjective opinions about each video.

We found that the screen size of standard cell phones today (approx. 2.2” x 1.6”) is not big enough for intelligible sign language, even at very high visual quality. Also, a larger screen size (4.5” x 3.1”) is not always as favorable as a medium screen size (3.1” x 2.0”); the medium size is comparable to the the biggest mobile phone displays available today. These are promising results for using existing technology for communication via sign language.

Neither screen size nor video distortion affected the visual significance of objects in the video (such as the signer’s face, arms, and torso). However, both screen size and video distortion *did* affect the amount of participant eye movement. The medium sized video (the one the participant’s favored) required the least amount of eye movements, with the larger and smaller sized videos both requiring much more. Furthermore, videos that were more highly distorted consistently required more participant eye movement to comprehend than less distorted videos.

Finally, we attempt to categorize visual excursions where the participant’s gaze-point falls far outside the signer’s head region for more than one second. Video content during these excursions include finger spelling, the signers hands moving to the bottom of the screen, the signer looking away from the camera, the signer pointing away from himself, the signer making lots of quick movements with his hands, as well as reasons unknown to the authors.

Results of this research could help define a new video compression metric which could then be used to create ASL-specific compression methods. For example, when considering region of interest encoding, the significance of any given region likely will not change for different screen sizes or levels of distortion. But, significance of regions *will* likely change if specific categories of sign mentioned above could be detected.

RELATED WORK

A good deal of research has been conducted with the aim of enhancing sign language communication with technology. Capture gloves have been used to sense various hand movements and hand shapes and then translate them to English [10, 19]. While this could be useful in a number of domains, the gloves are not able to sense contextual information such as signs that require both hands, signs that depend on hand placement in reference to the body, and the extremely crucial component of facial expression. There have been several attempts, including Eisenstein et al. [8] and Kadir et al. [9], to use computer vision and learning techniques to recognize hand gestures and facial expressions as signs and translate them to English. These are computationally expensive, but are getting more accurate. There have also been many attempts to render sign language animations from English text which have been useful in customer service, medical, and emergency situations [7, 12, 16, 20].

Rather than focus on ways to translate between sign language and spoken or written language, our focus is on using existing technology (mobile video phones) to facilitate remote communication between members of the Deaf Community. Our work is inspired by others who have studied eye movements of people watching sign language videos with implications for video compression. Muir et al. [11] found that viewers of sign tend to focus their gaze around the face and mouth of the signer with occasional excursions to the arms and torso. These findings seem to indicate that region of interest (ROI) encoding of video may be useful. ROI encoding has been investigated by Schumeyer et al. [15] amoung others. Unfortunately, these methods have not yet achieved real time performance.

In order for an ASL-specific compression method to be effective, we must understand in detail how ASL videos are viewed and understood by people fluent in ASL. In our work, we investigate how screen size and distortion due to compression affect how sign language videos are perceived.

EXPERIMENTAL DESIGN

The studies were performed in LUTE (Laboratory for Usability Testing and Evaluation) at the University of Washington. This laboratory has eye tracking capability through the ERICA system and GazeTracker software, both from Eye Response Technologies [4, 5]. The eye tracking method used in our study is the pupil center corneal reflection technique. The eye tracking system is accurate up to two degrees visual angle, which is 0.83” when the participant’s eye is 24” from the monitor. This particular eye tracking system tracks approximately 90% of the population.

We extracted video segments from *Signing Naturally Workbook and Video Text Expanded Edition: Level 1* [18]. Our segments varied in length from 7.2sec to 150.9sec, with a median length of 59.6sec and a mean length of 53.2sec. The x264 codec, an open source implementation of the H.264 (MPEG-4 part 10) codec, was used for compression [1, 2]. All videos were compressed at 29.97 frames per second.

We investigated several screen sizes and compression rates of videos. We chose these factors because cell phones have small screens [3, 13, 14, 17] and cell phone networks have slow transfer rates [6], making these two important factors for video playback on mobile phones. For each factor, we collected both qualitative and quantitative data. The following sections describe these four factors.

Video size

We studied three video sizes: large (4.5”x3.1”), medium (3.1”x2.1”) and small (2.2”x1.6”). The medium size is comparable to the largest screen size currently available on cell phones; the small size is comparable to a standard cell phone display. All videos were compressed at 96 kilobits per second (kbps).

Video compression rates

The rates we used were 16kbps, 24kbps, 48kbps, and 96kbps. 16kbps is comparable to maximum upload rates

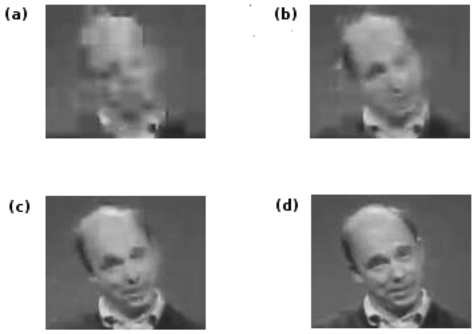

Figure 2. Cropped video frame at (a) 16kbps, (b) 24kbps, (c) 48kbps, and (d) 96kbps.

achievable on today's cell phone networks [6]. 96kbps is the highest rate at which our hardware could reliably playback video while recording eye movements. For all four compression rates, we set the video size to be large (4.5"x3.1"). Figure 2 gives an example of how the appearance varied at the different bit rates.

Qualitative data

We designed a five question, multiple choice survey which was given on the computer at the end of each video. The first question asked about video content, for example, "What was the name of the main character in the story?". The other four questions were repeated for each video:

- How difficult would you say it was to comprehend the video?
- During the video, about how often did you have to guess about what the signer was saying?
- How would you rate the annoyance level of the video?
- If video cell phones were available today with this quality of transmission, would you use them?

This post-video survey would pop up automatically once the video was completed. The answers were selected by mouse click on the appropriate choice.

Quantitative data

The eye tracking device recorded movements of the participants pupil as they watched the videos. For comparison, we collected movement data for hands and mouth in each video. This was done by running the videos at 30% speed and recording mouse location while an experimenter moved the mouse arrow to follow the left/right hand or head. This data was then normalized to original speed.

METHODOLOGY

Five members of the ASL community (2 women, 3 men) volunteered to participate in our study. Each study lasted approximately 45 minutes, with some participants completing in 25 minutes. Each study had two sections, one to study size and the other to study compression rate.

Participant Number	Age	Sex	Language Preference	Years ASL knowledge
1	51	M	ASL	51
2	46	M	English	27
3	21	F	ASL	21
4	35	M	ASL	35
5	20	F	English	8

Table 1. Participant demographics.

The size section collected qualitative data regarding the preferred screen size. The compression rate section collected qualitative data regarding the lowest compression rate acceptable to the participants. During both sections eye tracking data was recorded to investigate changes to eye movement patterns caused by video size or compression.

After entering the laboratory, each participant was told that he or she would watch several videos of various quality and size, while their eye movements were recorded with the ERICA eye tracking device. After signing an informed consent form, participants were given a demographic survey. They were then seated 24" away from the monitor and calibrated to the eye tracking device.

A practice video provided a brief introduction to the video watching system, ERICA eye tracking device and to the post-video survey. Then, each participant watched six videos which varied over the three chosen sizes with a 96kbps compression. Subsequently, the participant watched another 6 videos where the compression rate of the videos was varied over the three lower compression rates. Within each section, we varied the order in which participants were shown videos of different sizes or compression rates.

RESULTS

Demographics

Results from the demographic survey can be found in Table 1. Participant 2 was a hearing ASL-English translator; the other participants were deaf.

Participant 1 did not calibrate to the eye tracking device. This participant watched the videos and answered the post-video questions without recording any eye tracking data. In total, we received qualitative data from five participants and quantitative data from four.

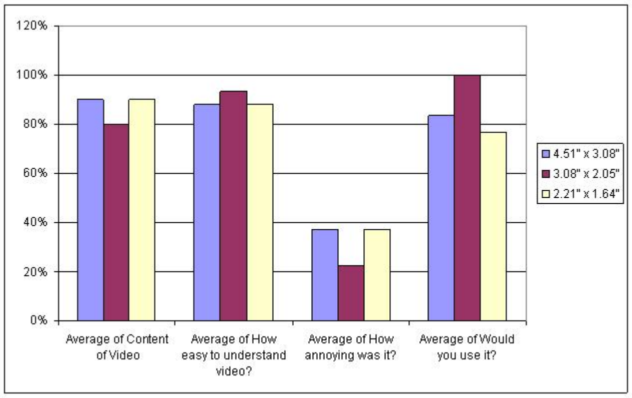

Figure 3. Qualitative results for different video sizes

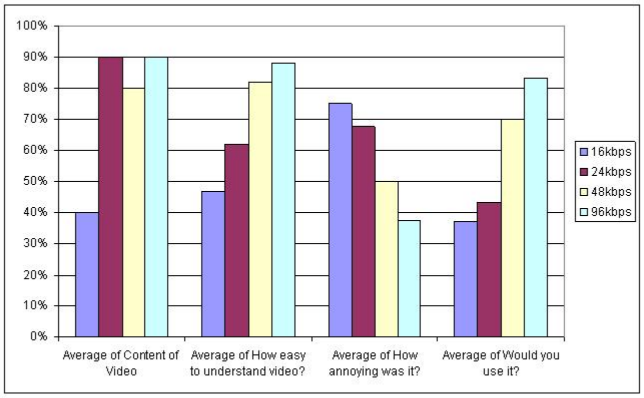

Figure 4. Qualitative results for different video compression rates

Qualitative Data

Qualitative data was obtained from the post-video questionnaires which asked participants subjective questions about the videos. Unfortunately, due to a reversal of our Likert scale, we had to throw out the third question which asked how often the participant guessed about the content of the video. Responses seemed to show half of the participants answering automatically instead of what they may have wished to answer. The answers to the questions are reproduced below. Notice the reverse ordering of positive to negative answers in the third question compared to the second or fourth.

Content

video specific

Difficulty

very difficult, difficult, neither, easy, very easy

Interpolation

not at all, some of the time, half of the time, most of the time, all of the time

Annoyance

very annoying, somewhat, a little, not annoying at all

Use

yes, maybe, no

Video Size

The perceived quality of large, medium and small video sizes was studied. Our goal was to work with the highest video quality possible for this part of the study. This goal was constrained by the eye tracking device itself. While recording, the eye tracking device uses a large amount of memory, which did not allow us to play videos at the highest DVD video quality. As such we chose the highest quality that ran smoothly, 96kbps, which was also the lowest quality at which the authors could not visually detect significant compression distortions.

Figure 3 shows answers to post-video surveys averaged over participants. For the four questions the y-axis represents the percentage of correct answers, perceived ease of video viewing, video annoyance, and the likeliness of mobile phone use with video size, respectively. The medium screen size was best received by the participants. The percentage of correct answers to the first question (video content) do not follow this trend; we speculate that this is due to the experimenter's inability to assess the difficulty of the questions. For example, we do not know how easy or difficult it would be to answer a question like “What was the name of the main character in the story?” given the video shown.

Many of our participants commented that the small sized videos were very difficult to watch. However, it is not the case that the bigger is always better, in fact most participants preferred the medium size. We believe that this is due to the fact that all videos in this part of the study were compressed at 96kbps. The two smaller videos, because they were smaller (with no other variables modified) were of better viewing quality than the largest video. This was also compounded by the memory use of the eye tracking software. Since the largest video size was our upper limit for frame rate, the video would occasionally skip, freeze and drop frames as the eye tracker memory use increased.

Video compression rates

The perceived quality of 16kbps, 24kbps, 48kbps, and 96kbps video compression rates was studied. All of the videos in this part of the study were viewed at the large video size. Since all of the videos in the first part of the study were compressed with 96kbps, the large size in the first part of the study served as the least compressed video in the second part of the study.

Figure 4 shows responses averaged over participants. For the four questions the y-axis represents the percentage of correct answers, perceived ease of video viewing, video annoyance, and the likeliness of mobile phone use with video size, re-

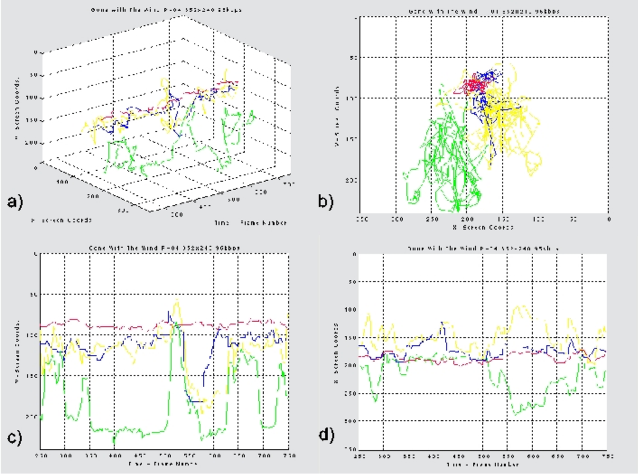

Figure 5. One participant's gaze trail (blue) in reference to the signer's mouth (red), left hand (green), and right (yellow) hands over the course of 500 frames of one video. (a) X and Y screen coordinates of the data trails for each frame over time, (b) data trail aggregated through time, (c) Y screen coordinates of data trails over time, and (d) X screen coordinates of data trails over time.

spectively. The highest compression rate was best received by the participants. There does seem to be a non-uniform gap between the 24kbps and the 48kbps which could indicate some sort of threshold acceptance rate. This would be an interesting area for future investigation.

Quantitative Data

While watching the videos, the participant's eye movements were tracked using an eye tracker. From this data we attempted to answer two main questions: (1) Does the size or the quality of the video affect where people focus their gaze and how often they move their gaze around? and (2) Given that previous research (discussed in Related Work) suggests the majority of gaze-points occur near the signer's head, what causes the occasional excursions to the arms and torso?

Figure 5 shows an example of a typical gaze trail with reference to the mouth, left hand, and right hand trails of the signers throughout 500 frames of one video clip.

Video size

Analysis of the eye movement data did in fact confirm earlier research by Muir et. al. [11]: nearly 95% of gaze points fell within 3 degrees of visual angle of the signer's mouth (at 24" away from the monitor, 3 degrees visual angle is approximately 1.25"). As the size of the video decreases, many more gaze points fall within 1 and 2 degrees visual angle, but this is simply because the content of the video (the signer's head and body) are also at a smaller visual angle. So, for the screen size analysis, we scaled the distance from the gaze to the head to be proportional to the size of objects in the video. For a gaze-point to be within 1 unit of the signer's head, it must be within 1 degree visual angle in the large

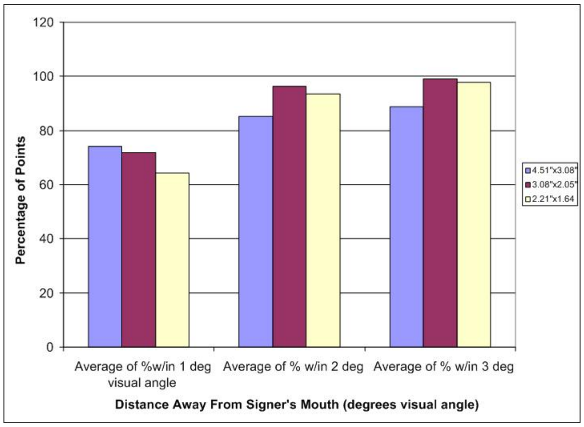

Figure 6. Percentage of gaze-points that fell within 1, 2, and 3 units from the signer's mouth, averaged over participants. There is not a significant difference when video size is varied.

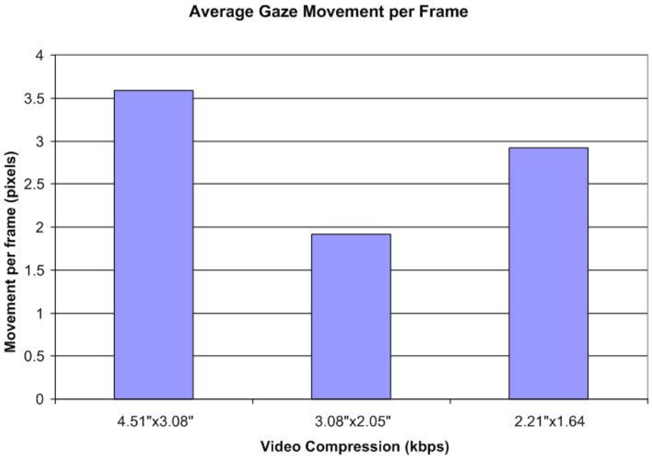

Figure 7. Distance per frame of each gaze-point to the next, averaged over participants. The medium sized videos required less eye movement than the large videos, which required even less movement than the smallest sized videos.

(4.51”x3.08”) sized videos, 2/3 degree visual angle in the medium (3.08”x2.05”) videos, and 1/2 degree visual angle in the small (2.21”x1.64”) videos.

Figure 6 shows that there is not a significant difference in where people focus their gaze when watching different sized sign language videos. This indicates that size does not have an affect on the visual significance of objects in the video.

This information does not indicate how much the participant is moving their gaze during the video. To discover this, we calculated the distance between each gaze point per frame. If the participant is moving their gaze around more frequently (imagine the gaze trail in Figure 5 becoming more squiggly), then the total distance covered by the trail will be greater than if the participant’s gaze trail were more smooth. The distance we calculate approximates the amount a participant’s eyes move during the video. Figure 7 shows the results of this analysis averaged over all participants. This indicates that the smallest sized videos required the most amount of eye movement and, interestingly, the medium sized videos required the least amount of movement. We speculate that this is the reason that the participants favored the medium size in the qualitative responses. Videos that require less eye movement to comprehend are favored over those which require more eye movement.

Video compression rates

An analysis similar to that above was conducted for videos of varying quality. Findings were similar when analyzing the number of gaze-points that fell with 1, 2, and 3 degrees visual angle from the head (Figure 8). There is no significant difference in the visual significance of objects in the video when the video is more or less distorted. This is important from a data compression standpoint, especially when considering region of interest coding, because an ASL-specific compression method will likely not need to modify the significance of each area when more or less distortion is required.

In order to answer the question of whether video quality affects the amount of eye movement needed to comprehend the video, we again calculated the distance per frame of the gaze trail and averaged over participants in Figure 9. We found that quality *does* affect how much the participant’s moved their eyes, moving around more for more distorted videos and less for less distorted videos. This correlates with the above hypothesis that videos requiring less eye movements are favored.

Excursions

In addition to analyzing where people focus their gaze and how much they move their eyes around, we also analyzed what causes visual excursions from the head and mouth of the signer. We looked at a sampling of gaze trail data from the videos that the participant’s watched and found occasions where the gaze was more than 3 degrees visual angle away from the signer’s mouth for at least 30 frames (approximately 1 second). We categorized these occasions into one of the following categories: fingerspelling, the signers hands

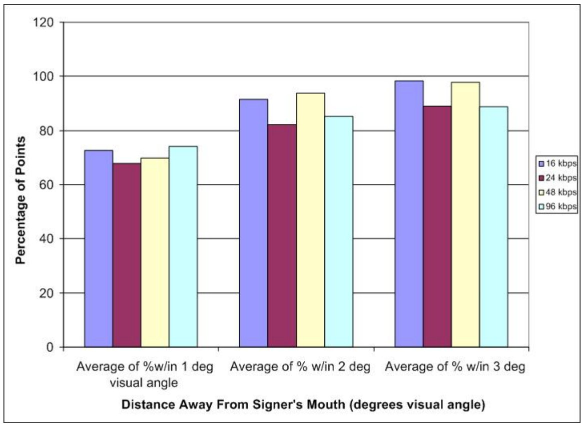

Figure 8. Percentage of gaze-points that fell within 1, 2, and 3 degrees visual angle from the signer’s mouth, averaged over participants. There is not a significant difference when bit rate is varied.

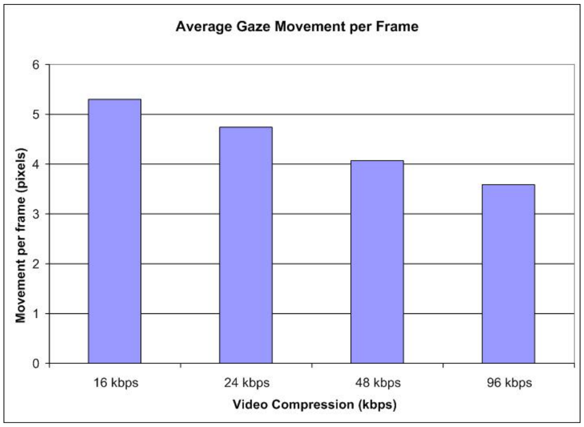

Figure 9. Distance per frame of each gaze-point to the next, averaged over participants. For more distorted (less bit rate) videos, more eye movements were required to comprehend the video.

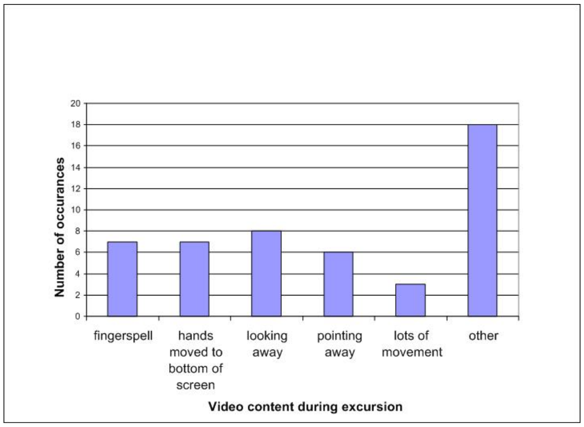

Figure 10. Number of excursions (where distance from gaze-point to signer's mouth were more than 3 degrees visual angle for a minimum of 1 second) in each category of video content.

moving to the bottom of the screen, the signer looking away from the camera, the signer pointing away from himself, the signer making lots of quick movements with his hands, and other causes. Figure 10 shows the number of each type of occasion and suggests that conversational content may result in visual excursions. We did find one instance of two participants making similar excursions during the same frames of the same videos. This implies that the event which caused the excursion could be something that generally triggers visual excursions of receivers of sign. Further analysis may result in very interesting information of this nature.

FUTURE WORK

Deeper analysis of the eye tracking data we gathered is necessary. In general, the participant's pupil is focused on the head of the speaker in the videos. As the focus strays from the head, it usually travels toward the hands. Further analysis of video contents is needed to determine the exact reasons for these excursions.

This study was small. A larger scale study would allow us to more accurately generalize the results. We are especially curious to see if there are any differences in the viewing styles of those for whom ASL is a primary language and those for whom it is a secondary language.

One participant informed us that an adult using ASL daily would not sign as slowly or precisely as the signers in our videos. Repeating the study with more realistic and less formal videos could provide useful insight.

We only studied the affect of size and compression rate on video viewing. Other technical factors that could affect the comprehensibility of the videos are frame rate and how often frames are dropped. Another factor could be the platform on which the videos are played. Our study was performed by playing videos on a computer monitor; perhaps results would be different if the videos were played directly on a mobile phone.

The long term goal of this project is to enable members of the Deaf Community to communicate using mobile video phones. Developing compression specific to sign language and testing it with users is important future work. Once a compression scheme is developed, it would be interesting to investigate domains other than mobile video phones where it could be used.

CONCLUSION

This study investigated the visual and perceptual effects of varying screen size and distortion (due to compression) of sign language videos. We found that screen size does affect the ease of use and enjoyment of the viewer, with today's largest available mobile phone display sometimes preferred over larger screen sizes if the smaller screen affords higher quality video. We also found that people visually attend to similar regions of the video regardless of screen size or distortion. Finally, videos that were preferred by participants were also the ones that required the least amount of eye movement: less distorted videos and videos of medium (3.1" x 2.0") size resulted in less total distance traveled by the eyes than other videos.

The results of this research demonstrate that there are several factors that contribute to the ability to comprehend sign language. They include, but certainly are not limited to, screen size, level of distortion, packet loss, formality of sign, video content, amount of signer movement, and size of signer movement.

These findings are significant from a video compression standpoint. Based on our results, a region of interest encoder would not need to redefine regions for videos of different size or visual quality. Our results indicate that existing mobile phone technology, when coupled with a new means of compression, could be suitable for sign language communication. This combination could provide access to the freedom, independence, and portable convenience of the wireless telephone network from which deaf Americans have previously been excluded.

REFERENCES

1. Richardson, I.E.G., "vcodex : H.264 tutorial white papers". `http://www.vcodex.com/h264.html`.

2. Aimar, L., Petit, E., Chen, M., Clay, J., Rullgård, M. and Merritt, L. VideoLAN - x264. `http://www.videolan.org/x264.html`

3. Cell Phones from Motorola: Wireless Phones, Motorola Mobile Phones. `http://direct.motorola.com`

4. Eye Response Technologies — ERT's ERICA eye gaze system. `http://www.eyeresponse.com/ericasystem.html`.

5. Eye Response Technologies — GazeTracker eye gaze analysis software. `http://www.eyeresponse.com/analysis.html`

6. General Packet Radio Service. `http://en.wikipedia.org/wiki/GPRS`

7. J.A. Bangham, S.J. Cox, M. Lincoln, I. Marshall, M. Tutt, and M Wells. *Signing for the deaf using virtual humans*. In IEE Colloquium on Speech and Language processing for Disabled and Elderly, 2000.

8. Eisenstein, J., Ghandeharizadeh, S., Huang L., Shahabi, C., Shanbhag, G. and Zimmermann, R., *Analysis of Clustering Techniques to Detect Hand Signs*, In Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing, 2001.

9. Kadir T., Bowden R., Ong E., Zisserman A. *Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition*. In Proc. BMVC04. Kingston UK. Sept 2004. Vol 2, pp939-948.

10. Kadous, M., *GRASP: Recognition of Australian sign language using instrumented gloves*, Honours thesis, School of Computer Science and Engineering, University of New South Wales, 1995.

11. L. Muir, I. Richardson, and S. Leaper. *Gaze tracking and its application to video coding for sign language*. Picture Coding Symposium, April 2003.

12. M. Verlinden, C. Tijsseling, H. Frowein (IvD, Sint-Michielsgestel), *Sign language on the WWW*, Proceedings of 18th International Symposium on Human Factors in Telecommunication (HFT2001), November, 2001

13. Nokia Cellular Phones. `http://www.nokiausa.com`.

14. palmOne - Products - Treo 650 Smartphone. `http://www.handspring.com/products/smartphones/treo650/index.jhtml`.

15. R. Schumeyer, E. Heredia, and K. Barner *Region of Interest Priority Coding for Sign Language Videoconferencing*. IEEE First Workshop on Multimedia Signal Processing, pp. 531-536, Princeton, 1997.

16. S. Augustine Su and Richard Furuta, *VRML-based Representations of ASL fingerspelling on the World-Wide Web*, ASSETS'98 - The Third International ACM SIGCAPH Conference on Assistive Technologies, April, 1998.

17. Samsung's Digital World - Products. `http://product.samsung.com`.

18. Smith, C., Mikos, K. and Lentz. E.M., *Signing Naturally Workbook and Videotext Expanded Edition: Level 1*. Dawnsign Press, San Diego, CA, USA, 1993.

19. Starner, T., J. Weaver, and A. Pentland, *Real-Time American Sign Language Recognition Using Desk and Wearable Computer-Based Video*, In IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 20 (12), pp. 1371.1375, December 1998.

20. VCom3D ASL Animations Signing Software `http://www.vcom3d.com/SignSmithDemo.htm`