MobileASL: Intelligibility of Sign Language Video as Constrained by Mobile Phone Technology

Anna Cavender, Richard E. Ladner
Department of Computer Science & Engineering
University of Washington
Seattle, Washington 98195 USA
{cavender, ladner}@cs.washington.edu

Eve A. Riskin
Department of Electrical Engineering
University of Washington
Seattle, Washington 98195 USA
riskin@ee.washington.edu

ABSTRACT

For Deaf people, access to the mobile telephone network in the United States is currently limited to text messaging, forcing communication in English as opposed to American Sign Language (ASL), the preferred language. Because ASL is a visual language, mobile video phones have the potential to give Deaf people access to real-time mobile communication in their preferred language.

However, even today's best video compression techniques can not yield intelligible ASL at limited cell phone network bandwidths. Motivated by this constraint, we conducted one focus group and one user study with members of the Deaf Community to determine the intelligibility effects of video compression techniques that exploit the visual nature of sign language. Inspired by eyetracking results that show high resolution foveal vision is maintained around the face, we studied region-of-interest encodings (where the face is encoded at higher quality) as well as reduced frame rates (where fewer, better quality, frames are displayed every second). At all bit rates studied here, participants preferred moderate quality increases in the face region, sacrificing quality in other regions. They also preferred slightly lower frame rates because they yield better quality frames for a fixed bit rate. These results show promise for real-time access to the current cell phone network through sign-language-specific encoding techniques.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues- Assistive technologies for persons with disabilities; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems- Video

General Terms

Design, Experimentation, Human Factors

ASSETS'06, October 22–25, 2006, Portland, Oregon, USA.

Keywords

Video Compression, American Sign Language (ASL), Deaf Community, Mobile Telephone Use

1. INTRODUCTION

MobileASL is a video compression project that seeks to enable wireless cell phone communication through sign language.

1.1 Motivation

Mobile phones with video cameras and the ability to transmit and play videos are rapidly becoming popular and more widely available. Their presence in the marketplace could give Deaf [1] people access to the portable conveniences of the wireless telephone network.

The ability to wirelessly transmit video, as opposed to just text or symbols, would provide the most efficient and personal means of mobile communication for members of the Deaf Community: deaf people, family members, and friends who use ASL. Some members of the Deaf Community currently use text messaging (such as SMS, IM, or TTY), but text is cumbersome and impersonal because (a) English is not the native language of most Deaf people in the United States (ASL is their preferred language), and (b) text messaging is slow and tedious at 5-25 words per minute [9] compared to 120-200 wpm for both spoken and signed languages. Many people in the Deaf Community use video phones which can be used to call someone with a similar device directly or a video relay service (where a remote human interpreter translates video sign language to spoken language). 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 [10]. Video cell phones have the potential to make the mobile phone network more accessible to over one million Deaf people in the U.S. [11].

Unfortunately, the Deaf Community in the U.S. cannot yet take advantage of this new technology. Our preliminary studies strongly suggest that even today's best video encoders cannot produce the quality video needed for intelligible ASL in real time, given the bandwidth and computational constraints of even the best video cell phones.

Realistic bit rates on existing GPRS networks typically vary from 30-80kbps for download and perhaps half that for

[1]Capitalized Deaf refers to people who are active in the signing Deaf Community and Deaf Culture, whereas lowercase deaf is typically a more medical term.

upload [7]. While the upcoming 3G standard [2] and special rate multi-slot connections [7] may offer much higher wireless bit rates, video compression of ASL conversations will still play an important role in realizing mobile video phone calls. First, there is some uncertainty about when 3G technology will become broadly available and, when it does, it will likely be initially restricted to highly populated areas and suffer from dropped calls and very poor quality video as is currently the case in London [1]. Furthermore, degradations in signal-to-noise-ratio conditions and channel congestion will often result in lower actual bit rates, packet loss, and dropped calls. More importantly, fair access to the cell phone network means utilizing the already existing network such that Deaf people can make a mobile video call just as a hearing person could make a mobile voice call: without special accommodations, more expensive bandwidth packages, or additional geographic limitations. As such, video compression is a necessity for lowering required data rates and allowing more users to operate in the network, even in wireless networks of the future. The goal of the MobileASL project is to provide intelligible compressed ASL video, including detailed facial expressions and accurate movement and gesture reproduction, at less than 30 kbps so that it can be transmitted on the current GPRS network. A crucial first step is to gather information about the ways in which people view sign language videos.

1.2 Contributions

We conducted one focus group and one user study with local members of the Deaf Community in Seattle to investigate the desire and/or need for mobile video phone communication, the technical and non-technical challenges involved with such technology, and what features of compressed video might enhance understandability.

The purpose of the focus group was to elicit feedback from the target user group about the ways in which mobile video communication could be useful and practical in their daily lives.

The user study was inspired by strongly correlated eye movement data found independently by Muir *et al.* [12] and Agrafiotis *et al.* [3]. Both projects used an eyetracker to collect eye movement data while members of the Deaf Community watched sign language videos. Results indicate that over 95% of gaze points fell within 2 degrees visual angle of the signer's face. The reasoning is that skilled receivers of sign focus their gaze on or near the lower face of the signer. This is because contextual information coming from the hands and arms is relatively easy to perceive in peripheral or parafoveal vision, whereas contextual information from the face of the signer (which is also important in sign language) requires the level of detail afforded by high resolution foveal vision.

Based on these results, we conducted a study to investigate effects of two simple compression techniques on the intelligibility of sign language videos. We created a fixed region-of-interest (ROI) encoding by varying the level of distortion in a fixed region surrounding the face of the signer. The ROI encodings result in better quality around the face at the expense of quality in other areas. We also varied the frame rates of the video so that for a given bit rate, either 15 lower quality frames or 10 higher quality frames were displayed per second.

Results from the study indicate that minor adjustments to standard video encoding techniques, such as ROI and reduced frame rates, may allow intelligible ASL conversations to be transmitted in real-time over the current U.S. cell phone network.

We will next discuss related work in Section 2. In Section 3, we will share participant responses from the MobileASL focus group. Section 4 explains the video compression user study. Section 5 discusses the results of the study, and Section 6 presents future work and concludes.

2. RELATED WORK

Previous research has studied the eye movement patterns of people as they view sign language through the use of an eyetracker ([12], [3]). Both groups independently confirmed that facial regions of sign language videos are perceived at high visual resolution and that movements of the hands and arms are generally perceived with lower resolution parafoveal vision. Both groups recommend a video compression scheme that takes these visual patterns into account (such as a region-of-interest encoding). Video segmentation of the important regions of sign language videos has been implemented using several different methods and shows promise for reducing bit rate through region-of-interest encodings ([8], [14]). None of these methods have been empirically validated by potential users of the system.

Furthermore, guidelines recommending between 12 and 20 frames per second have been proposed for low bit rate video communication of sign language [15]. These claims have also not been empirically validated to the authors' knowledge.

Another line of research has pursued the use of technology to interpret between spoken languages and signed languages (for example [17], [16], [5]). While these translation technologies may become useful in limited domains, the goal of our project does not involve translation or interpretation.

Rather than focusing on ways to translate between written/spoken and signed languages, we feel that the best way to give Deaf people access to the conveniences of mobile communication is to bring together existing technology (such as large screen mobile video phones) with existing social networks (such as ASL interpreting services). The only missing link in this chain of communication is a way to transfer intelligible sign language video over the mobile telephone network in real time.

3. FOCUS GROUP

We wanted to learn more about potential users of video cell phone technology and their impressions about how, when, where, and for what purposes video cell phones might be used. We conducted a one-hour focus group with four members of the Deaf Community ranging in age from mid-twenties to mid-forties. The conversation was interpreted for the hearing researcher by a certified sign language interpreter. The discussion centered around the following general topics and responses are summarized below:

Physical Setup

The camera and the screen should face the same direction. Most current phones have cameras facing away from the screen so that one can see the picture while aiming the camera. This obviously would not work for filming oneself, as in a sign language conversation.

The phone should have a way to prop itself up, such as

a kickstand. Participants agreed that some conversations could occur while holding the phone and signing at the same time. But, for longer conversations, it would be desirable to put the phone on a table or shelf.

The phone should be slim, not bulky, so that it could be carried in a pocket or purse. However, connecting a camera that captures better quality video should be an option (this concept is similar to using a Bluetooth headset).

The phone should have a full keyboard, like the PDA-style phones, not just numbers 0-9, as text will likely still be an important means of communication.

Features

All participants agreed that the phone should have all of the features currently found in Sidekicks or Blackberry PDA-phones, such as email and instant messaging. People in the Deaf Community have become accustomed to having these services and won't want to carry around two separate devices.

Even though participants all agreed that video communication is a huge improvement over text, they still felt that text messages would be an important feature to have. For example, text could be used to initiate a phone call (like ringing someone), troubleshoot (e.g. "I can't see you because..."), or simply as a fall back mechanism when the connection is bad or when conditions or situations aren't favorable for a video call. All participants thought that text should also be an option during a video call, much like the simultaneous text messaging options in online video games.

The phone should alert the user of requests for a video call just as typical phones do for voice calls. There should be an easy way to accept or decline a video call. When a call is declined, the caller should be able to leave a video message. Participants decided to call this SignMail as opposed to VoiceMail.

Video calls should be accessible to/from other video conferencing software so that calls can be made between video cell phones and web cams or set top boxes.

Packet Loss

Networks are notoriously unreliable and information occasionally gets lost or dropped. The solution to this in a video sign language conversation is simply to ask the signer to repeat what was missed. However, all participants agreed that video services would not be used, or paid for, if packet losses were too frequent.

Privacy Concerns

We thought that perhaps using a video phone to communicate might involve privacy concerns if other people, especially those who know sign, can see the screen of the phone. However, this turned out to be a non-issue. Because of the visual nature of sign language, anyone with in view can eavesdrop on the conversation, whether face-to-face or over video. The response was simply, "If the conversation is private, take it somewhere private."

Scenarios

We also discussed several scenarios where the video phone might or might not be useful. Two examples of these scenarios are as follows:

What if the phone rings when driving or on the bus?

There should be an easy way to dismiss the call, or change the camera angle so that the phone could be placed in one's lap while on the bus. The phone could also be mounted on the dash board of a car. People already sign while driving, even to people in the back seat through the rear-view mirror, so this wouldn't be very different. It could be as dangerous as talking while on the cell phone and participants thought its use may be affected by future cell phone laws.

What if there were no table available to set the phone down?

Signing with one hand for short conversations would not be a problem. People sign while drinking, eating, smoking, etc. But, if the location is bad, like a crowded bar, texting might be easier.

One participant succinctly explained, "I don't foresee any limitations. I would use the phone anywhere: at the grocery store, on the bus, in the car, at a restaurant, on the toilet, anywhere!"

In order for these scenarios to become reality, a better method for encoding (and compressing) video is needed such that intelligible ASL can be transmitted over the low bandwidth cell phone network.

4. VIDEO COMPRESSION STUDY

Inspired by the results from Muir *et al.* [12] and Agrafiotis *et al.* [3], we conducted a study with members of the Deaf Community to investigate the intelligibility effects of (a) three levels of increased visual clarity in a small region around the face of the signer (region-of-interest, or ROI) as well as (b) two different frame rates (frames per second or fps). These factors were studied at three different bit rates comparable to those available in the current U.S. cell phone network, totaling 18 different encoding techniques. Eighteen different sign language videos were created for this study so that each participant could be exposed to every encoding technique without watching the same video twice (i.e. a repeated measures design).

The videos were recordings of short stories told by a local Deaf woman at her own natural signing pace. They varied in length from 0:58 minutes to 2:57 minutes (mean = 1:58) and all were recorded with the same location, lighting conditions, background, and clothing. The x264 codec, an open source implementation of the H.264 (MPEG-4 part 10) codec, was used to compress the videos with the 18 encoding techniques [4, 13]. See Appendix A for a complete listing of encoding parameters used for the study videos.

Both videos and questionnaires were shown on a Sprint PPC 6700, PDA-style video phone with a 320×240 pixel resolution (2.8" x 2.1") screen. All studies were conducted in the same room with the same lighting conditions.

4.1 Baseline Video Rating

Original recordings yielded 22 total videos of which 18 were chosen for this study for the following reasons. Undistorted versions of all 22 videos were initially rated for level of difficulty by 3 separate participants (1 Deaf, 2 hearing) who considered themselves fluent in ASL. The purpose of the rating was to help eliminate intelligibility factors not related to compression techniques. After viewing each video, participants were asked one multiple choice question about

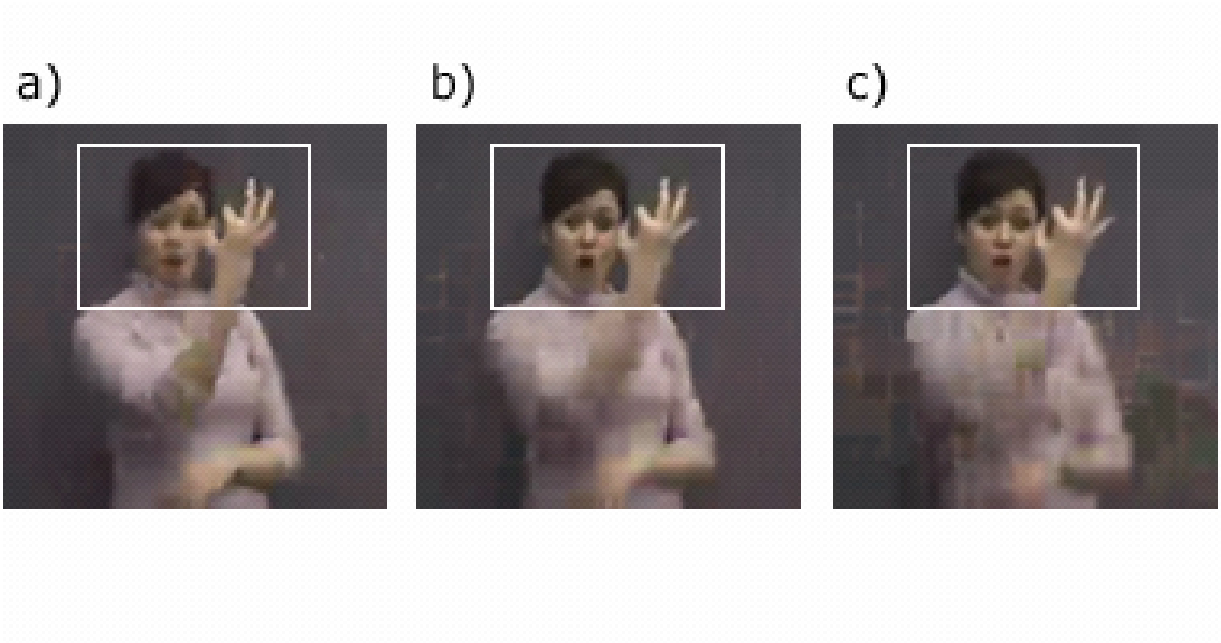

Figure 1: Cropped video frame at (a) -0 ROI (standard encoding), (b) -6 ROI (two times better quality in the face region), and (c) -12 ROI (four times better quality in the face region).

the content of the video and then asked to rate the intelligibility of the video using a 5-point Likert scale with unmarked bubbles on a range from "difficult" to "easy." We will refer to those bubbles as "1" through "5" here.

The first participant rated all 22 videos as "5," the second rated twenty of the videos as "5" and two as "4," and the third participant also rated twenty of the videos as "5" and two as "4" (although the two were distinct from the ones rated a "4" by the second participant). The four videos that were given a rating a "4" were excluded from the study so that only the remaining 18 videos were used. In fact, post hoc analysis of the results from the study found no significant differences between the ratings of any of these 18 videos. This means we can safely assume that the intelligibility results that follow are due to varied compression techniques rather than other confounding factors (e.g. signer speed, difficulty of signs, lighting or clothing issues that might have made some videos more or less intelligible than others).

4.2 Bit rates

We studied three different bit rates: 15, 20, and 25 kilobits per second (kbps). We chose these values in an attempt to accurately portray the current U.S. mobile phone network: the optimal download rate has been estimated at 30 kbps whereas the upload rate is considerably less, perhaps as low as 15 kbps.

4.3 Frame rates

We studied two different frame rates: 10 and 15 frames per second (fps). Preliminary tests with a certified sign language interpreter revealed that 10 fps and 15 fps were both acceptable for intelligible ASL. The difference between 30 fps and 15 fps was negligible whereas at 5 fps signs became difficult to watch and fingerspelling became nearly impossible to understand.

Frame rates of 10 and 15 fps were chosen for this study to investigate the tradeoff of fewer frames at slightly better quality or more frames at slightly worse quality for any given bit rate. For example, a video encoded at 10 fps has fewer frames to encode than the same video at 15 fps, so more bits can be allocated to each frame.

4.4 Region of Interest

Finally, we studied three different region-of-interest (ROI) values: -0, -6, -12, where the negative value represents the reduced quantizer step size, out of 52 possible step sizes, in a fixed 6×10 macroblock region around the face (a single 320×240 pixel frame is composed of 15×20 macroblocks). Reducing the quantizer step size in this region results in less compression (better quality) in the face region and more compression (sacrificing quality) in all other regions for a given bitrate. An ROI value of -0 means there is no difference in the aforementioned regions (i.e. a typical encoding). An ROI value of -6 doubles the quality in the face region, distributing the remaining bits over the other regions. And an ROI value of -12 results in a level of quality four times better than a typical encoding around the signer's face sacrificing even more quality in surrounding regions. The three ROI values shown to participants can be seen in Figure 1.

As with frame rate, the ROI values for this study were chosen based on preliminary studies conducted with a certified sign language interpreter.

4.5 Video Order

Because we can assume that higher bit rates yield more intelligible videos, we chose to structure the order in which videos were shown so that analysis of the data for the three bit rates could safely be separated. Thus, the study was partitioned into three parts, one for each bit rate. The same videos were shown in each partition, but their order within

the partition was randomized. The order with which the three parts of study were conducted was determined by a Latin-squares design. The order with which the six different encodings (combinations of 2 frame rates and 3 ROIs) were shown within each part was also determined by a Latin-squares design (meaning each participant watched all 18 different encodings in a different order to avoid effects of learning and/or fatigue).

4.6 Subjective Questionnaire

After each video, participants answered a three-question, multiple choice survey given on the phone's screen and answered using the phone's stylus (see Figure 2(a)-(c)). The first question asked about the video content, for example, "Who was the main character in the story?" This question was simply asked in order to encourage participants to pay close attention to the content of the videos, not necessarily to assess their understanding of the video. It would have been extremely difficult to devise questions of similar linguistic difficulty across all videos. Although we were not interested in the answers to this first question, it is worth mentioning that the correctness of the participants' answers often did not correlate with their answers to the remaining questions. For example, it was not uncommon for the participants to answer the first question correctly and then report that the video was difficult to comprehend and that they would not use a mobile phone that offered that quality of video, and vice versa.

The remaining two questions were repeated for each video. These two questions appear in Figure 2 (b) and (c) and ask (1)"How easy or how difficult was it to understand the video?" which we will refer to as "understand" and (2) "If video of this quality was available on a cell phone, would you use it?" which we will refer to as "use." Answers to both questions were constrained by a 5-point Likert scale (just as in the Baseline Video Rating) where participants could choose from five bubbles labeled with the ranges (1) difficult ... easy and (2) no ... maybe ... yes.

5. RESULTS OF VIDEO PHONE STUDY

Eighteen adult members of the Deaf Community (7 women, 11 men) participated in the study. 10 participants were Deaf, 5 were hearing, and 3 were CODAs (Child of a Deaf Adult). CODAs are born into Deaf households and often consider ASL their first language and English their second language. All 10 Deaf participants and 3 CODAs had lifelong experience with ASL; the 5 hearing participants had an average of 10.6 years (SD = 5.81) experience with ASL.

After a short demographic survey, participants were shown two practice videos. These videos also served as examples of the worst and best videos to be viewed so that the participants had points of reference when rating upcoming videos.

Participants then watched 18 videos and answered the three-question, multiple-choice survey for each video. Each video was encoded with a distinct encoding technique as described above. The last 5 to 10 minutes of each study session was spent gathering anecdotal information about the participant's impressions of the video cell phone and video quality.

Analysis of survey questions responses indicates that participant preferences for all three variables (bit rate, frame rate, and region-of-interest or ROI) were largely independent of each other. For example, the results for ROI encodings held true regardless of changes in bit rate or frame rate.

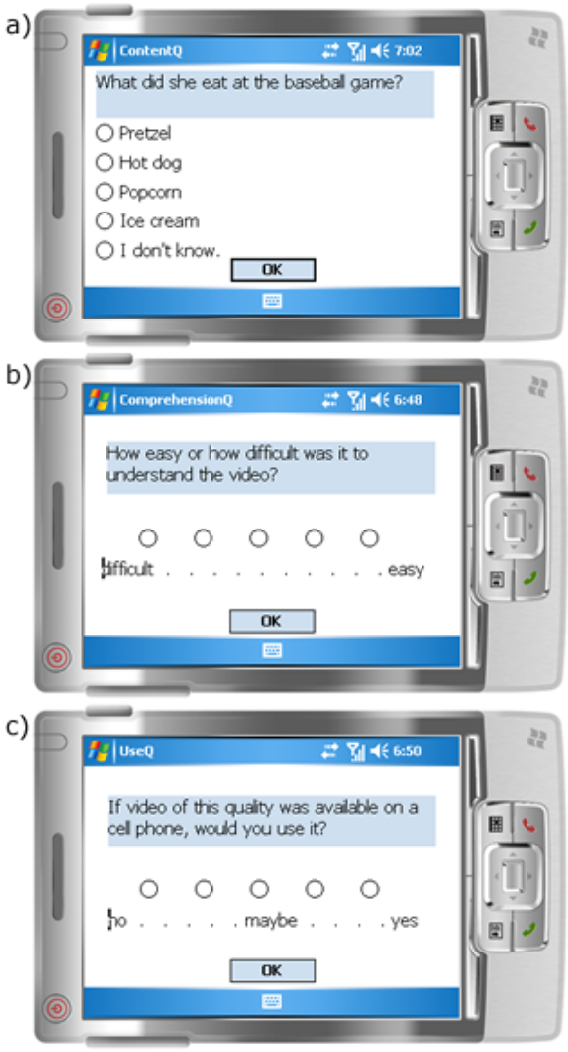

Figure 2: Series of questions in post-video questionnaires asking about (a) the content of the video (different for each video), (b) the understandability of the video, and (c) the participant's willingness to use the quality of video just seen.

Also, answers to the two questions "understand" and "use" were highly correlated (with a Pearson's Correlation Coefficient $r(16) = .85$ and $p < .01$). Because "understand" and "use" are so highly correlated, we will only show graphs for "understand" below, but note that data for "use" looks very similar.

5.1 Bit rates

As expected, survey responses indicate very strong and statistically significant preferences for higher bit rates: 25 kbps was preferred over 20 kbps, which in turn was preferred over 15 kbps ($F(2, 34) = 51.12$, $p < .01$). These results can be seen in Figure 3.

Higher bit rates were preferred regardless of different frame rates and region-of-interest (ROI) values of the videos.

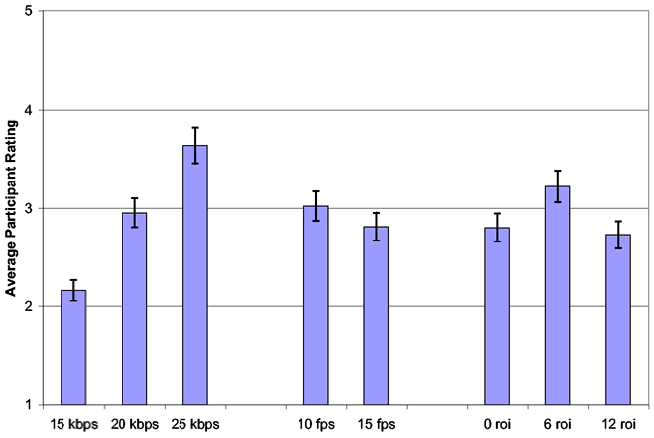

Figure 3: Qualitative results for different bit rates, frame rates, and region-of-interest values, averaged over participants. Error bars represent confidence intervals.

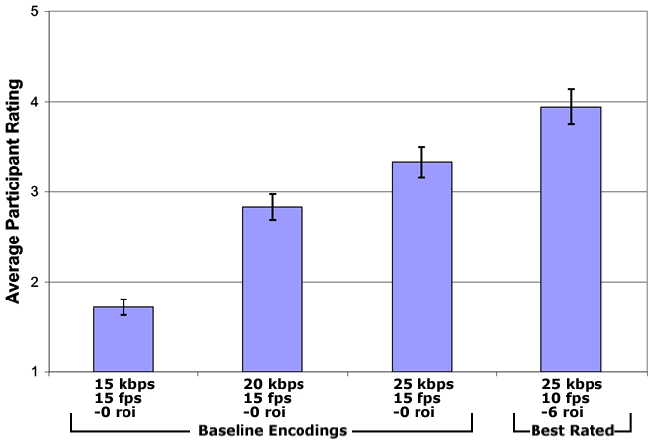

Figure 4: Qualitative results for the three baseline bit rate encodings (25, 20, and 15 kbps at 15 fps, -0 roi) shown against the encoding technique with the maximum average rating (25 kbps, 10 fps, -6 roi). Error bars represent confidence intervals.

5.2 Frame rates

For the two different frame rates studied here (10 fps and 15 fps), Figure 3 shows a preference toward 10 fps ($F(1, 17) = 4.59$, $p < .05$). A likely explanation is that the difference between 10 fps and 15 fps is acceptable (in fact some participants did not realize the videos had different frame rates until after the study, during the anecdotal questions). Furthermore, the increased frame quality, due to the fewer number of frames to encode, may have been a desirable tradeoff.

5.3 Region of Interest

For the three different region-of-interest encodings, we found a very significant effect ($F(2, 34) = 13.69$, $p < .01$). At every bit rate shown, participants preferred an ROI of -6 (the middle value tested). Figure 3 shows that ratings for -0 ROI (no region of interest) and -12 ROI were not statistically different. This could indicate that an appropriate range of values was chosen for the study and that there may exist an optimal tradeoff between clarity around the face and distortion in other regions.

Figure 4 shows the average participant ratings for the baseline encoding techniques for this study (i.e. 15 frames per second and no region-of-interest) at the three bit rates tested. The graph also shows the encoding technique that received the best average participant rating: 25 kbps, 10 fps, and -6 roi. There is a substantial increase in preference for encodings that add both a reduced frame rate and a moderate region-of-interest improvement.

These results indicate that these two simple and effective compression techniques may better utilize low bandwidth connections (such as through the cell phone network) to yield more intelligible ASL. Also, reducing the frame rate to 10 fps and decreasing the quantization by 6 step sizes in macroblocks around the face can be executed without increases in encoding time (which is important when encoding on the limited processors of cell phones).

6. CONCLUSION

The Deaf Community in the United States stands to benefit from new video cell phone technology as a mobile form of communication in their preferred language, American Sign Language (ASL). Low bandwidth constraints of the current cell phone network create many video compression challenges because even today's best video encoders cannot yield intelligible ASL at such low bit rates. Thus, real-time encoding techniques targeted toward sign language are needed to meet these demands.

This paper discussed the potential for and challenges involved with mobile video sign language communication over the U.S. cell phone network. We investigated the potential needs and desires for mobile sign language communication through a targeted focus group with four members of the Deaf Community.

Motivated by highly correlated visual patterns of receivers of sign found by Muir *et al.* [12] and Agrafiotis *et al.* [3], we studied the effects of a region-of-interest encoding (where the face regions of the video were encoded at better quality than other regions) and reduced frame rate encodings (where fewer, better quality frames are displayed every second). We studied these two factors at three different bit rates representing a lower-end possible range for transfer over the current cell phone network. Results indicate that reducing the frame rate to 10 frames per second (fps) and increasing the quality of the image near the signer's face may help yield more intelligible ASL for low bit rates. Increased quality per frame for 10 fps was a preferable tradeoff from 15 fps. A tradeoff of 6 decreased quantization steps near the face of the signer (doubling the quality in that region) was preferred over a typical (no region-of-interest) encoding and over a larger quality increase in the face that caused more distortion in other regions of the video (a quantization difference of 12).

These findings are important from a video compression standpoint. 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 people in the United States have previously been excluded.

6.1 Future Work

The results from this work are currently being used to help define a new video compression metric that will inform a compression scheme utilizing the new H.264 standard [6]. The following are future goals we have for this project.

The regions used in this study for the region-of-interest (ROI) encodings were fixed in size and location. While the signer in our videos did not move her upper body outside of this rectangle, and since most signs occur within a "sign box" surrounding the upper body, it may be more useful and/or more efficient to dynamically choose a region of interest based on information in the video. Many participants thought that the ROI was most problematic when the shape of the hands was lost due to distortion. An interframe skin detection algorithm [8] could likely help with this as it may include regions associated with hands in the ROI encoding. Similarly, motion vectors in the video (supposing a stationary background) could help define regions of interest where more bits could be allocated. This would be a valuable idea to test empirically as a moving region of interest may be distracting.

Because of the natural constraints of sign language (the shape, orientation, and location of arms, hands, and face) it may be useful to apply learning algorithms to the motion vectors of several training videos so that the motion vectors in other sign language videos may be more easily predicted, aiding in the speed and efficiency of the compression. For example, an encoder that could predict times of signing, fingerspelling, or "just watching" could act accordingly: saving bits when no signing is occurring or allocating more bits when high quality video is needed, such as during the highly detailed movements of fingerspelling.

Participants in the Focus Group all agreed that packet loss will be a big concern and could render mobile video technology useless for ASL conversations. An area of future research will be investigating the implications, limitations, and effective ways to handle packet loss for video sign language communication.

The long term goal of this project is to enable members of the Deaf Community to communicate using mobile video phones. Developing compression techniques that encode and decode in real-time on a mobile phone processor is an important and on-going aspect of this project. For example, we are working on an encoder with a constant encoding time, which will adjust parameters in the coder based on how long it is taking to encode. Specifically, an encoder could reduce time spent searching for good motion vectors when decreased encoding time is needed whereas improvements to the encoding can be made when more encoding time is available. We will continue to optimize the H.264 encoder for both speed and video quality.

Acknowledgements

We would like to thank Sheila Hemami, Francis Ciaramello, Alan Borning, and Rahul Vanam for their guidance and feedback throughout this project.

We would also like to thank Janet Davis, Kasia Wilamowska, and Stefan Schoenmackers for help with editing various versions of this paper and Kate Deibel for help with the statistical design and analysis of the study.

Finally, thanks to Jessica DeWitt for helping to create the videos used in this study, Tobias Cullins for arranging interpreters, and all of the people who participated in the study.

This research has been supported by the National Science Foundation through two Grants CCF-0514353 and CCF-0104800, an NSF Graduate Fellowship, and a Boeing Professorship.

7. REFERENCES

[1] 3GNewsroom.com. 3UK disgraced by BBC watchdog programme. http://www.3gnewsroom.com/, October 22, 2003.

[2] 3GToday. http://www.3gtoday.com/, 2006.

[3] D. Agrafiotis, C. N. Canagarajah, D. R. Bull, M. Dye, H. Twyford, J. Kyle, and J. T. Chung-How. Optimized sign language video coding based on eye-tracking analysis. In *VCIP*, pages 1244–1252, 2003.

[4] L. Aimar, L. Merritt, E. Petit, M. Chen, J. Clay, M. Rullgrd, C. Heine, and A. Izvorski. x264 - a free h264/AVC encoder. http://www.videolan.org/x264.html, 2005.

[5] J. 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.

[6] F. Ciaramello, A. Cavender, S. Hemami, E. Riskin, and R. Ladner. Predicting intelligibility of compressed american sign language video with objective quality metrics. In *2006 International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2006.

[7] GSMA. General packet radio service. http://www.gsmworld.com/technology/gprs/class.shtml, 2006.

[8] N. Habili, C.-C. Lim, and A. Moini. Segmentation of the face and hands in sign language video sequences using color and motion cues. *IEEE Trans. Circuits Syst. Video Techn.*, 14(8):1086–1097, 2004.

[9] C. L. James and K. M. Reischel. Text input for mobile devices: comparing model prediction to actual performance. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 365–371, 2001.

[10] E. Keating and G. Mirus. American sign language in virtual space: Interactions between deaf users of computer-mediated video communication and the impact of technology on language practices. In *Language in Society*, volume 32, pages 693–714, 2003.

[11] R. Mitchell. How many deaf people are there in the United States? http://gri.gallaudet.edu/Demographics/deaf-US.php, 2005.

[12] L. Muir and I. Richardson. Perception of sign language and its application to visual communications for deaf people. In *Journal of Deaf Studies and Deaf Education*, volume 10, pages 390–401, 2005.

[13] I. Richardson. vocdex : H.264 tutorial white papers. http://www.vcodex.com/h264.html, 2004.

[14] R. Schumeyer, E. Heredia, and K. Barner. Region of Interest Priority Coding for Sign Language Videoconferencing. In *IEEE First Workshop on Multimedia Signal Processing*, pages 531–536, 1997.

[15] I. T. S. Sector. Draft application profile: Sign language and lip reading real time conversation usage of low bit rate video communication. 1998.

[16] T. Starner, A. Pentland, and J. Weaver. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1371–1375, 1998.

[17] C. Vogler and D. Metaxas. A Framework for Recognizing the Simultaneous Aspects of American Sign Language. *Comput. Vis. Image Underst.*, 81(3):358–384, 2001.

APPENDIX

A. X264 PARAMETERS USED IN VIDEO PHONE STUDY

The following list the parameters used for the x264 encodings of the video used the Video Phone Study discussed in this paper.

resolution: 320x240

–subme 6 (Sub-pixel motion estimation, partition decision quality: 6=best)

–bframes 1 (Number of B-frames between I and P, default is 0)

–no-b-adapt (Disable adaptive B-frame decision)

–scenecut -1 (How aggressively to insert extra I-frames, default=40)

-I 9999 (Maximum GOP size, default is 250)

–mixed-refs (Decide references on a per partition basis)

–me umh (Pixel motion estimation method, uneven multi-hexagon search)

–direct spatial (Direct MV prediction mode, spatial)

–ref 5 (Number of reference frames, default is 1)

-A p8x8,p4x4,b8x8,i8x8,i4x4 (Partitions to consider during analysis)