MobileASL: Overcoming the technical challenges of mobile video conversation in sign language

Anna C. Cavender[1], Neva Cherniavsky[1], Jaehong Chon[2], Richard E. Ladner[1],
Eve A. Riskin[2], Rahul Vanam[2], Jacob O. Wobbrock[3]

[1]Computer Science & Engineering, [2]Electrical Engineering, [3]The Information School
University of Washington
Seattle, WA USA 98195
[1]*(cavender, nchernia, ladner)@cs.washington.edu*
[2]*(jaehong, riskin, rahulv)@ee.washington.edu*
[3]*wobbrock@u.washington.edu*

Abstract

As part of the ongoing MobileASL project, we have built a system to compress, transmit, and decode sign language video in real-time on an off-the-shelf mobile phone. In this work, we review the challenges that arose in developing our system and the algorithms we implemented to address them. Separate parts of this research have been previously published in (Cavender et al., 2006; Cherniavsky et al., 2007; Cherniavsky et al., 2008; Vanam et al., 2009; Chon et al., 2009; Cherniavsky et al., 2009).

Compression and transmission of sign language video presents unique difficulties. We must overcome weak processing power, limited bandwidth capacity, and low battery life. We also must ensure that the system is usable; that is, that the video is intelligible and the algorithms that we employ to save system resources do not irritate users.

We describe the evolution of the MobileASL system and the algorithms we utilize to achieve real-time video communication on mobile phones. We first review our initial user studies to test feasibility and interest in video sign language on mobile phones. We then detail our three main challenges and solutions. To address weak processing power, we optimize the encoder to work on mobile phones, adapting a fast algorithm for distortion-complexity optimization to choose the best parameters. To overcome limited bandwidth capacity, we utilize a dynamic skin-based region of interest, which encodes the face and hands at a higher bit rate at the expense of the rest of the image. To save battery life, we automatically detect periods of signing and lower the frame rate when the user is not signing.

We implement our system on off-the-shelf mobile phones and validate it through a user study. Fluent ASL signers participate in unconstrained conversations over the phones in a laboratory setting. They find the conversations with the dynamic skin-based region of interest more intelligible. The variable frame rate affects conversations negatively, but does not affect the users perceived desire for the technology.

Ongoing work includes varying the spatial resolution instead of the temporal resolution, further optimization of rate-distortion-complexity, and a field study to determine usability over a long period of time in a realistic setting.

1. Introduction

Mobile technology has become an integral part of society, changing the nature of communication worldwide. The MobileASL project aims to expand accessibility for Deaf[1] people by efficiently compressing sign language video to enable mobile phone communication. Users capture and receive video on a typical mobile phone. They wear no special clothing or equipment, since this would make the technology less accessible.

There are three main challenges to building a system for real-time two-way video communication on mobile phones. First, the processing power on phones is weak. The encoder must run fast enough to show the video in real-time, and yet must produce intelligible video at low bit rates. Secondly, the bandwidth is limited. Video must be transmitted at rates of less than 30 kbps to be compatible with the capacity of the U.S. mobile phone network. Lastly, the battery capacity is low. Encoding, transmitting, receiving, and playing video on a mobile phone quickly drains the battery, rendering the phone useless.

We develop sign language sensitive algorithms to attack these three challenges. We optimize the encoder parameters for the best possible tradeoff between efficiency and intelligibility, using an adaptation of a fast algorithm for distortion-complexity optimization. We address the problem of limited bandwidth by creating a dynamic skin-based *region-of-interest* (ROI) that encodes the face and hands at a higher bit rate at the expense of the rest of the image, increasing intelligibility without increasing bandwidth. We save power and processor cycles through automatic detection of periods of signing. When the user is not signing, we lower the frame rate, encoding and transmitting one tenth of the frames. We call this technique *variable frame rate* (VFR).

Our central goal is to increase access for Deaf people; we thus use intelligibility as our main measure of success. Throughout the evolution of our system, we verify our design and algorithms with users. We began the project by conducting focus groups and small laboratory studies to

[1] Capitalized Deaf refers to members of the signing Deaf community, whereas deaf is a medical term.

validate our ideas. After building a working system, we evaluate it with a larger study in which fluent signers participate in unconstrained conversations over the phone.

1.1. Background

As is often the case with the design and implementation of a large system, separate parts of this research have been published previously (Cavender et al., 2006; Cherniavsky et al., 2007; Cherniavsky et al., 2008; Vanam et al., 2009; Chon et al., 2009; Cherniavsky et al., 2009). More complete versions of related work may be found in those publications.

Sign language video compression so that Deaf users can communicate over the telephone lines has been studied since at least the early 1980s. The first works attempted to enable communication by drastically modifying the video signal, e.g. by binarizing the image; (Foulds, 2006) provides a good overview. More closely related to our project are works that implement ROI encoding for reducing the bit rate of sign language video (Schumeyer et al., 1997; Woelders et al., 1997; Saxe and Foulds, 2002; Agrafiotis et al., 2003; Habili et al., 2004) and works that examine the intelligibility of sign language video at low frame rates (Sperling et al., 1986; Parish et al., 1990; Johnson and Caird, 1996; Hooper et al., 2007). Most of the ROI algorithms were not evaluated with Deaf users and are not real-time. Research into low frame rates for sign language are inconsistent in their conclusions, but there appears to be a sharp drop off in intelligibility at frame rates lower than 10 *frames per second* (fps).

MobileASL is built on top of the latest standard in video compression, H.264 (Wiegand et al., 2003). The H.264 encoder works by dividing a frame into 16 × 16 pixel *macroblocks*. It compares each macroblock to those sent in previous frames, looking for exact or close matches. The macroblock is then coded with the location of the match, the displacement, and whatever residual information is necessary. We use the Open Source x264 (Aimar et al., 2005; Merritt and Vanam, 2007) codec.

2. Design of the MobileASL System

The design of the MobileASL system is closely based on the needs and desires of users, and informed by a focus group and user studies.

2.1. Focus group

In our initial focus group, we find that users want a "smart" phone that has a front-side camera, a full keyboard, full email and instant messaging abilities, and a kick stand so that the phone can be placed on the table. Users also want to be able to use the phones to access video relay services, which allow communication between Deaf and hearing via sign language interpreters, and to chat with other users who have web cams or set top boxes. Based on these results, we choose to use HTC TyTN-II smart phones running Windows Mobile 6.1 (Qualcomm MSM7200, 400 MHz ARM processor, Li-polymer battery). The video size is QCIF (176 × 144). Figure 1 shows a phone running MobileASL. Our system is not currently able to handle calls to other devices, but we hope to add that functionality in the future.

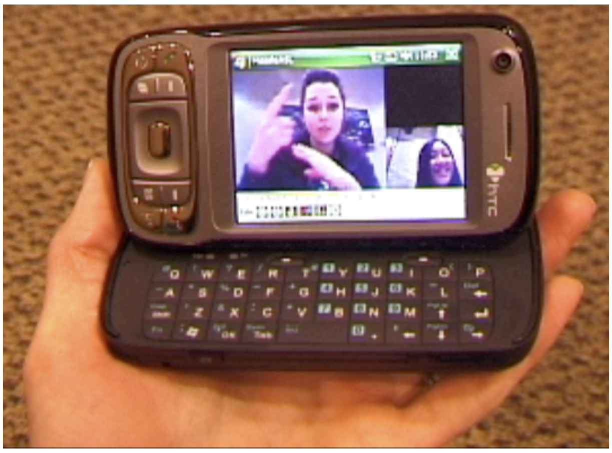

Figure 1: MobileASL running on the HTC TyTN-II

2.2. Initial ROI and VFR evaluation

In several initial user studies, we investigate the feasibility of our ROI and VFR techniques. We find that videos with ROI are intelligible, up to a point; however, when too many bits are devoted to the face at the expense of the rest of the frame, it becomes distracting for users. For the variable frame rate, users evaluate conversational sign language videos that have (artificially created) lower frame rates during periods of not signing. We find that users dislike an entirely frozen frame for the not signing portions, but otherwise rate the quality similarly. As there is no large drop off in the perception of intelligibility, we use both methods in our system.

3. Sign language sensitive compression

To address the three main challenges of weak processing speed, limited bandwidth, and low battery life, we implement the following techniques for sign language sensitive video compression: optimal parameter selection for encoder optimization, dynamic skin-based ROI, and variable frame rate.

3.1. Optimal parameter selection

The H.264 encoder has many different parameters that are possible to tune to achieve the highest quality possible video at the lowest possible cost. For example, there are several different methods for searching the macroblocks for matching, with varying complexity. However, it is computationally infeasible to test all possible combinations of parameter settings for a given bit rate. Using a variation of the GBFOS (Chou et al., 1989) and ROPA (Kiang et al., 1992) algorithms, we jointly optimize H.264 encoder parameter settings for quality and complexity. We are able to search through many fewer encodings to arrive at the optimal selection.

3.2. Dynamic skin-based ROI

Given the parameter settings, H.264 will try to encode the frame at the highest possible quality for the bit rate. One way to increase intelligibility while maintaining the same bit rate is to shift the bits around, so that more are focused on the face and less are focused on the background. Using a simple range query on the chrominance components, we determine the macroblocks that contain a majority of skin pixels, and encode these at a higher quality setting (allocating more bits to the important part of the frame). Since the encoder is constrained by the bit rate, the result is that the other macroblocks in the frame are encoded with fewer bits and correspondingly lower quality.

3.3. Variable frame rate

Sign language video is conversational and involves turn-taking, meaning that often when one person is signing, the other person is not. We aim to automatically recognize when a user is not signing and lower the frame rate from 10 fps to 1 fps. Since far fewer frames are encoded and transmitted, this results in a large power savings, allowing conversations to go on much longer. We obtain a power gain of 8% over the battery life of the phone, corresponding to an extra 23 minutes of talk time.

Automatic recognition on the phone is challenging for the same reasons as the overall system implementation. We must be able to perform the recognition in real-time while hopefully not adding to the complexity. To this end, we use a simple differencing method to distinguish signing frames from not signing frames. The sum of absolute differences of the luminance component is calculated between successive raw frames and compared to a previously determined threshold. This is temporally smoothed by applying a sliding window that takes the average vote over the window and classifies the frame accordingly. The average classification accuracy as measured on a frame-by-frame basis on videos taken with the phone camera is 76.6%.

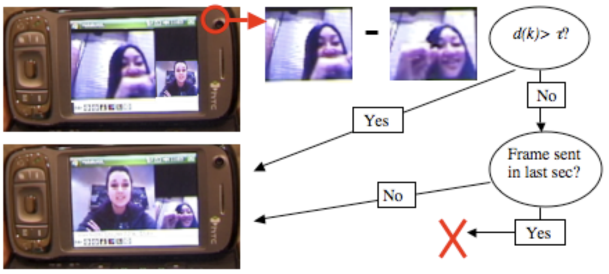

Figure 2: The architecture of the variable frame rate. Differences between frames are checked; if the user isn't signing, the frame is sent only to maintain 1 fps.

4. Evaluation

To validate our algorithms and test our working system, we conducted a user study with members of the signing Deaf community. Fifteen participants fluent in sign language took part in the study. For each conversation, participants sat on the same side of a table, separated by a screen, with a black background behind them (see Figure 3). Since we expect that Deaf people will use the phones in a variety of situations, we did not control for the relationship between participants. There were conversations between interpreters and native signers, between strangers and friends, and even between a married couple.

All combinations of three versions of ROI (no, low, and high) and two versions of VFR (off and on) were tested, for a total of six different possible settings. After five minutes of unconstrained conversation, the participants filled out a subjective questionnaire about their experience. They then continued their conversation under different settings. The order in which the settings were evaluated differed between users. Both sides of the conversations were captured by a third video camera, in order to obtain objective measures, such as the number of times a user asked for a repetition.

Figure 3: Study setting. The participants sat on the same side of a table, with the phones in front of them.

We statistically analyzed both the subjective and objective results of the user study. For the subjective measures, we found statistically significant differences in the perception of the number of guesses and comprehension. Using a high level of ROI decreased the number of guesses and increased comprehension. ROI did not statistically significantly affect the objective measures, but VFR did. The users asked for repeats more often and had more conversational breakdowns when the VFR was on than when it was off. This is probably due to classification inaccuracy resulting in mistakenly lowering the frame rate when the person is actually signing. Despite these measurable difficulties with VFR, there was no statistically significant difference in subjective measures for VFR; in particular, the users' perceived desire for the technology was unaffected. We expect that VFR is a feature that users will choose to employ depending on their needs, for example, if they are going on a trip and want to preserve battery life.

5. Future directions

In the future, we will continue to improve MobileASL so that we may make it widely available. Our next step is to move out of the lab and into the field. We plan to give participants phones with MobileASL installed and have them use and comment on the technology over an extended period of time.

Technically speaking, several challenges remain. We can improve classification accuracy by using more advanced machine learning techniques on the phone. We found in our user study that often our algorithm misclassified finger spelling frames, since users slowed down during those periods. If our classifier recognized finger spelling in addition to signing and not signing, we could adjust the frame rate accordingly. We also want to investigate different methods for saving power on the phone, such as changing the spatial resolution during not signing periods instead of lowering the frame rate. Furthermore, there is a continual trade-off in our system between the complexity of our algorithms, the speed at which we can encode, the intelligibility of the video, and the bit rate. We want to further explore jointly optimizing these conditions, ideally in real-time and as circumstances differ. For example, the encoder often struggles in noisy environments where there is a lot of background motion; in order to keep sending the frames in real time, we can reduce the quality, readjusting the parameters when circumstances improve.

The first question asked by users at the end of our study was always "when will this be available?" During the recruitment process, we received interested queries from all over the United States. Our ultimate goal is to make our technology widely available, so that Deaf people will have full access to today's mobile telecommunication network.

6. References

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

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

A. Cavender, R. E. Ladner, and E. A. Riskin. 2006. MobileASL: Intelligibility of sign language video as constrained by mobile phone technology. In *ASSETS '06: Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 71–78.

N. Cherniavsky, A. C. Cavender, R. E. Ladner, and E. A. Riskin. 2007. Variable frame rate for low power mobile sign language communication. In *ASSETS '07: Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*. ACM Press.

N. Cherniavsky, R. E. Ladner, and E. A. Riskin. 2008. Activity detection in conversational sign language video for mobile telecommunication. In *Proceedings of the 8th international IEEE conference on Automatic Face and Gesture Recognition*. IEEE Computer Society.

Neva Cherniavsky, Jaehong Chon, Jacob O. Wobbrock, Richard E. Ladner, and Eve A. Riskin. 2009. Activity analysis enabling real-time video communication on mobile phones for deaf users. In *UIST '09: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, Victoria, BC, Canada, October 4-7, 2009*, pages 79–88. ACM Press.

Jaehong Chon, Neva Cherniavsky, Eve A. Riskin, and Richard E. Ladner. 2009. Enabling access through real-time sign language communication over cell phones. In *43rd Annual Asilomar Conference on Signals, Systems, and Computers*. IEEE Computer Society.

Philip A. Chou, Tom D. Lookabaugh, and Robert M. Gray. 1989. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transactions on Information Theory*, 35(2):299–315.

R. A. Foulds. 2006. Piecewise parametric interpolation for temporal compression of multijoint movement trajectories. *IEEE Transactions on information technology in biomedicine*, 10(1):199–206.

N. Habili, C.-C. Lim, and A. Moini. 2004. Segmentation of the face and hands in sign language video sequences using color and motion cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(8):1086–1097.

S. Hooper, C. Miller, S. Rose, and G. Veletsianos. 2007. The effects of digital video quality on learner comprehension in an American Sign Language assessment environment. *Sign Language Studies*, 8(1):42–58.

B. F. Johnson and J. K. Caird. 1996. The effect of frame rate and video information redundancy on the perceptual learning of American Sign Language gestures. In *CHI '96: Conference companion on Human factors in computing systems*, pages 121–122. ACM Press.

S. Z. Kiang, R. L. Baker, G. J. Sullivan, and C. Y. Chiu. 1992. Recursive optimal pruning with applications to tree structured vector quantizers. *IEEE Transactions on Image Processing*, 1(2):162–169.

L. Merritt and R. Vanam. 2007. Improved rate control and motion estimation for H.264 encoder. In *ICIP '07: Proceedings of the 2007 IEEE International Conference on Image Processing*, volume 5, pages 309–312. IEEE Computer Society.

D. H. Parish, G. Sperling, and M. S. Landy. 1990. Intelligent temporal subsampling of American Sign Language using event boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):282–294.

D. M. Saxe and R. A. Foulds. 2002. Robust region of interest coding for improved sign language telecommunication. *IEEE Transactions on Information Technology in Biomedicine*, 6:310–316.

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

G. Sperling, M. Landy, Y. Cohen, and M. Pavel. 1986. Intelligible encoding of ASL image sequences at extremely low information rates. In *Papers from the second workshop Vol. 13 on Human and Machine Vision II*, pages 256–312, San Diego, CA, USA. Academic Press Professional, Inc.

Rahul Vanam, Eve A. Riskin, and Richard E. Ladner. 2009. H.264/MPEG-4 AVC encoder parameter selection algorithms for complexity distortion tradeoff. In *DCC '09: Proceedings of Data Compression Conference*, pages 372–381. IEEE Computer Society.

T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra. 2003. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576.

W. W. Woelders, H. W. Frowein, J. Nielsen, P. Questa, and G. Sandini. 1997. New developments in low-bit rate videotelephony for people who are deaf. *Journal of Speech, Language, and Hearing Research*, 40:1425–1433.