Activity Analysis Enabling Real-Time Video Communication on Mobile Phones for Deaf Users

Neva Cherniavsky,[1] *Jaehong Chon,*[2] *Jacob O. Wobbrock,*[3] *Richard E. Ladner*[1], *Eve A. Riskin*[2]

[1]Computer Science & Engineering
University of Washington
Box 352350
Seattle, WA USA 98195

[2]Electrical Engineering
University of Washington
Box 352500
Seattle, WA USA 98195

[3]The Information School
DUB Group
University of Washington
Box 352840
Seattle, WA USA 98195

ABSTRACT

We describe our system called *MobileASL* for real-time video communication on the current U.S. mobile phone network. The goal of MobileASL is to enable Deaf people to communicate with Sign Language over mobile phones by compressing and transmitting sign language video in real-time on an off-the-shelf mobile phone, which has a weak processor, uses limited bandwidth, and has little battery capacity. We develop several H.264-compliant algorithms to save system resources while maintaining ASL intelligibility by focusing on the important segments of the video. We employ a *dynamic skin-based region-of-interest (ROI)* that encodes the skin at higher quality at the expense of the rest of the video. We also automatically recognize periods of signing versus not signing and raise and lower the frame rate accordingly, a technique we call *variable frame rate (VFR)*.

We show that our variable frame rate technique results in a 47% gain in battery life on the phone, corresponding to an extra 68 minutes of talk time. We also evaluate our system in a user study. Participants fluent in ASL engage in unconstrained conversations over mobile phones in a laboratory setting. We find that the ROI increases intelligibility and decreases guessing. VFR increases the need for signs to be repeated and the number of conversational breakdowns, but does not affect the users' perception of adopting the technology. These results show that our sign language sensitive algorithms can save considerable resources without sacrificing intelligibility.

ACM Classification: H5.2 [**Information interfaces and presentation**]:Multimedia Information Systems–*Video*. K.4.2 [**Computers and Society**]: Social Issues–*Assistive technologies for persons with disabilities*.

General terms: Human Factors

Keywords: MobileASL, sign language, video compression, computer vision, region-of-interest, variable frame rate.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

UIST'09, October 4–7, 2009, Victoria, British Columbia, Canada..
Copyright 2009 ACM 978-1-60558-745-5/09/10 ...$10.00.

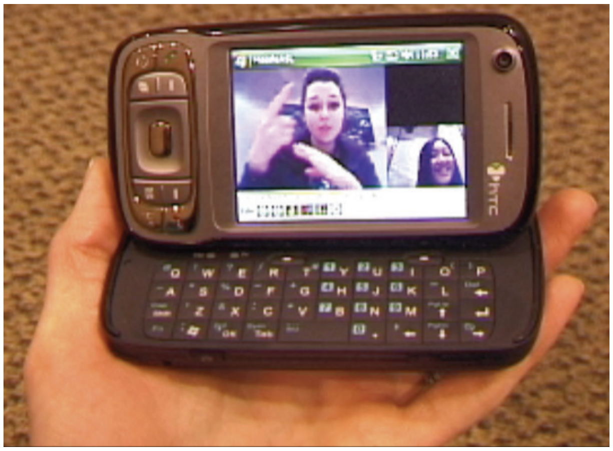

Figure 1: Deaf people can use MobileASL to communicate via real-time video on an off-the-shelf mobile phone over current non-3G cell phone networks.

INTRODUCTION

Mobile phone use has skyrocketed in recent years, with more than 2.68 billion subscribers worldwide [13]. Video mobile phones make it possible for Deaf[1] people to communicate in their native sign language. The explosion of mobile technologies has not left Deaf people behind; on the contrary, many regularly use mobile text devices such as Blackberries and Sidekicks. However, text messaging is much slower than signing. Signing has the same communication rate as spoken language of 120-200 words per minute (wpm) versus 5-60 wpm for text messaging [5, 14]. Furthermore, text messaging forces Deaf users to communicate in English as opposed to ASL.

The goal of *MobileASL* (Figure 1) is to provide real-time sign language video communication on off-the-shelf mobile phones between users that wear no special clothing or equipment. The challenges are three-fold:

- **Low bandwidth:** In the United States, the majority of the mobile phone networks uses *general packet radio service (GPRS)* [8], which can support bandwidth up to around 30-50 kbps [7]. Japan and Europe use the higher

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

bandwidth 3G network [12]. While real-time mobile video communication is already available there, the quality is poor, the videos are jerky, and there is significant delay.

- **Low processing speed:** Even the best mobile phones available on the market running an operating system like Windows Mobile and able to execute many different programs have very limited processing power. Our current MobileASL phones (HTC TyTN II) have a 400 MHz processor, versus 2.5 GHz or higher for a typical desktop computer. The processor must be able to encode and transmit the video in close to real-time; otherwise, a delay is introduced that negatively affects intelligibility.

- **Limited battery life:** A major side effect of the intensive processing involved in video compression on mobile phones is battery drain. Insufficient battery life of a mobile device limits its usefulness if a conversation cannot last for more than a few minutes. In an evaluation of the power consumption of a handheld computer, Viredaz and Wallach found that decoding and playing a video was so computationally expensive that it reduced the battery lifetime from 40 hours to 2.5 hours [29]. For a sign language conversation, not only do we want to play video, but we also want to capture, encode, transmit, receive and decode video, all in real-time. Power is in some ways the most intractable problem; while bandwidth and processing speed can be expected to grow over the next few years, battery storage capacity has not kept up with Moore's law.

MobileASL must overcome these challenges while producing intelligible sign language video. With intelligibility as the goal, we can harness the natural structure of two-sided conversation as well as linguistic aspects of sign language to save resources. We save processor cycles and power by utilizing a *variable frame rate (VFR)*. We automatically determine when the user is signing and encode and transmit at the highest possible frame rate. When the user is not signing, we lower the frame rate to 1 frame per second (*fps*).

By focusing on the important parts of the video, we try to increase intelligibility while maintaining the level of compression, thus addressing the challenge of low bandwidth. Given that much of the grammar of sign language is found in the face [27], we encode the skin at higher quality at the expense of the rest of the frame.

This paper describes our MobileASL system and details our implementation of sign language-sensitive algorithms for variable frame rate and dynamic skin-based region-of-interest bit allocation. We implement both of these features in the video encoder on the phone to enable real-time compression and transmission. We report on the classification accuracy of our variable frame rate and the power savings. Use of the variable frame rate results in a 47% power gain for the battery life of the phone, equivalent to 68 minutes of talk time.

We evaluate our system in a user study in which the participants carry on unconstrained conversation on the phones in a laboratory setting. We gather both subjective and objective measures from the users. The results of our study show that our skin-based *region-of-interest (ROI)* technique reduces guessing and increases comprehension. The VFR technique results in more repeats and clarifications and in more conversational breakdowns, but this does not affect participants' perceived likelihood of using the phone. Thus we can significantly decrease resource use and this does not appear to detract from the users' overwhelmingly favorable impression of the technology. Our techniques employed in MobileASL may also be useful in other work that incorporates real-time video in user interfaces, especially on mobile devices.

RELATED WORK

Most computer science research in assistive technology for the Deaf focuses on sign language recognition, in which researchers attempt to translate sign language into English text. Ong and Ranganath describe the state-of-the-art [19]. However, the goal of our project does not involve translation or interpretation. We focus instead on providing the same access to the mobile telecommunications network that hearing people enjoy.

Early Work in Sign Language Video Compression

Compression of sign language video so that Deaf users can communicate over the telephone lines has been studied since at least the early 1980s. The specified bandwidth of the copper lines that carry the voice signal is 9.6 kbps or 3 kHz, too low for even the best video compression methods 40 years later. The earliest projects compressed sign language video by reducing multi-tone video to a series of binary images and transmitting them; see [6] for an overview. This approach achieves very low bit rate but suffers from several drawbacks. First, the binary images have to be transmitted separately and compressed using runtime coding or other algorithms associated with fax machines. The temporal advantage of video, namely that an image is not likely to differ very much from its predecessor, is lost. Moreover, complex backgrounds will make the images very noisy, since the edge detectors will capture color intensity differences in the background; the problem only worsens when the background is dynamic. Finally, much of the grammar of sign language is in the face. In these projects, the facial expression of the signer is lost. The majority of the papers [6] have very little in the way of evaluation, testing the systems in an ad-hoc manner and often only testing the accuracy of recognizing individual signs.

Recent Work in Video Compression

With the advent of the Internet and higher bandwidth connections, researchers began focusing on compressing sign language video instead of an altered signal. One obvious way to compress video is to separately compress each frame, using information found only within that frame. This method is called *intra-frame coding*. However, as noted above, this negates the temporal advantage of video. Modern video compression algorithms use information from other frames to code the current one; this is called *inter-frame coding*. The latest standard in video compression is H.264. It performs significantly better than its predecessors, achieving the same quality at up to half the bit rate [30]. H.264 works by dividing a frame into 16×16 pixel *macroblocks*. These are compared to previously sent reference frames. The algorithm looks for exact or close matches for each macroblock from the reference frames. Depending on how close the match is, the macroblock is coded with the location of the match, the displacement via a *motion vector*, and whatever residual error information is necessary. Macroblocks can be subdivided to the 4×4 pixel level. When a match cannot be found, the macroblock is coded as an *intra-block*, only from information within the current frame.

Region-of-interest and foveal compression. The availability of higher quality video at a lower bit rate led researchers to explore modifying standard video compression to work well on sign language video. Many researchers were motivated by work investigating the focal region of ASL signers. Some research used an eye-tracker to follow the visual patterns of signers watching sign language video and determined that users focused almost entirely on the face [1, 18]. In some sense, this is intuitive, because humans perceive motion using their peripheral vision [3]. Signers can recognize the overall motion of the hands and process its contribution to the sign without shifting their gaze, allowing them to focus on the finer points of grammar found in the face.

One natural inclination is to increase the quality of the face in the video. Agrafiotis et al. [1] implemented *foveal* compression, in which the macroblocks at the center of the user's focus are coded at the highest quality and with the most bits; the quality falls off in concentric circles. Their videos were not evaluated by Deaf users. Similarly, Woelders et al. [31] took video with a specialized foveal camera and tested various spatial and temporal resolutions. Signed sentences were understood at rates greater than 90%, though they did not test the foveal camera against a standard camera.

As we have done in this work, other researchers have implemented region-of-interest encoding for reducing the bit rate of sign language video. A *region-of-interest*, or ROI, is simply an area of the frame that is coded at a higher quality at the expense of the rest of the frame. Schumeyer et al. [25] suggest coding the skin as a region-of-interest for sign language videoconferencing. Similarly, Saxe and Foulds [24] used a sophisticated skin histogram technique to segment the skin in the video and compress it at higher quality. Habili et al. [9] also used advanced techniques to segment the skin. None of these projects evaluated their videos with Deaf users for intelligibility, and none of the methods are real-time, making them unsuitable for our purposes.

Temporal compression. The above research focused on changing the spatial resolution to better compress the video. Another possibility is to reduce the temporal resolution. The temporal resolution, or frame rate, is the rate at which frames are displayed to the user. Early work [11, 21] found a sharp drop in intelligibility of sign language video at 5 fps. Parish and Sperling [20] created artificially subsampled videos with very low frame rates and found that when the frames are chosen intelligently (i.e., to correspond to the beginning and ending of signs), the low frame rate was far more understandable. Johnson and Caird [15] trained sign language novices to recognize 10 isolated signs, either as points of light or conventional video. They found that users could learn signs at frame rates as low as 1 fps, though they needed more attempts than at a higher frame rates. Sperling et al. [26] explored the intelligibility of isolated signs at varying frame rates. They found nonsignificant differences from 30 to 15 fps, a slight decrease in intelligibility from 15 to 10 fps, and a large decrease in intelligibility from 10 fps to 5 fps.

More recently, Hooper et al. [10] looked at the effect of frame rates on the ability of sign language students to understand ASL conversation. They found that comprehension increased from 6 fps to 12 fps and again from 12 fps to 18 fps. The frame rate was particularly important when the grammar of the conversation was more complex, as when it included classifiers and transitions as opposed to just isolated signs. Woelders et al. [31] studied both spatial resolution and temporal resolution and found a significant drop in understanding at 10 fps. At rates of 15 fps, video comprehension was almost as good as the original 25 fps video. Finger spelling was not affected by the frame rates between 10 and 25 fps, possibly because the average speed of finger spelling is five to seven letters per second and thus 10 fps is sufficient [22].

DESCRIPTION OF MOBILEASL

The MobileASL implementation is based on the Open Source x264 H.264 codec [2, 16]. The x264 encoder was compared with the JM reference encoder (ver 10.2) [17] and was shown to be 50 times faster, while providing bit rates within 5% for the same peak signal-to-noise ratio (PSNR) [16]. This makes it a good choice for H.264 video compression on low-power devices.

We use HTC TyTN-II phones (Windows Mobile 6.1, Qualcomm MSM7200, 400 MHz ARM processor, Li-polymer battery), chosen because they have a front camera on the same side as the screen (Figure 1). The video size is QCIF (176×144). There are two ways to increase the processing speed of compression. We perform assembly optimization using the ARMv6 single instruction multiple data assembly set, and convert the most computationally intensive operations, such as motion estimation, into assembly. We also use the lowest possible x264 settings, changing the code when necessary. Even with these settings, our phones are only able to encode at a maximum rate of 7-8 fps; the bottleneck in this case is not the bandwidth, but the processor. We therefore modify the encoder to include our VFR and ROI encoding.

Variable Frame Rate

Our goal is to determine if the user is signing or not in order to adjust the frame rate and save power. In previous research associated with MobileASL projects [4], researchers used advanced feature extraction techniques and machine learning to recognize signing and not signing periods on conversational web cam videos. They also incorporated features from both sides of the conversation to increase classification accuracy. Building on this work, we implement several different techniques for automatic recognition on the phones.

Baseline differencing. As a simple, baseline method, we calculate the sum of absolute differences between successive frames. Let $I_k(i,j)$ be the luminence component of pixel (i,j) in frame k. Then the sum of absolute differences for frame k is:

$$d(k) = \sum_{(i,j)\in I(k)} |I_k(i,j) - I_{k-1}(i,j)| \tag{1}$$

We check this value against a previously determined threshold τ arrived at by training on conversational sign language video. If $d(k) > \tau$, we classify the frame as signing.

Joint differencing with linear programming. Previously, researchers utilized differencing information from both sides of the phone conversation to increase classification accuracy [4]. The parameters were determined empirically. We improve this by posing the problem as a linear program. Define $d_1(k)$ as the sum of absolute differences from the primary side of the conversation and $d_2(k)$ as the sum of absolute differences from the secondary side of the conversation. We want to choose α, β, and τ such that the following is true for as many k as possible:

$$\begin{aligned} &\alpha d_1(k) - \beta d_2(k) > \tau && \text{when } k \text{ is a signing frame,} \\ &\alpha d_1(k) - \beta d_2(k) \leq \tau && \text{when } k \text{ is not.} \end{aligned}$$

The intuition for these equations is that the motion, and therefore the differences, will be high on one side and low on the other, indicating signing.

Let $C = \{c_1, ..., c_n\}$ be a vector of indicator variables where 1 indicates signing and -1 indicates not signing, $c_k \in \{-1, 1\}$. Then:

$$\begin{aligned} \alpha d_1(k) - \beta d_2(k) >&\quad \tau \quad \forall k | c_k = 1 \\ \alpha d_1(k) - \beta d_2(k) \leq&\quad \tau \quad \forall k | c_k = -1 \end{aligned}$$

Assume τ is positive. Let $\mu = \alpha/\tau$ and $\gamma = \beta/\tau$. Write:

$$\begin{aligned} \mu d_1(k) - \gamma d_2(k) >&\quad 1 \quad \forall k | c_k = 1 \\ \mu d_1(k) - \gamma d_2(k) \leq&\quad 1 \quad \forall k | c_k = -1 \end{aligned}$$

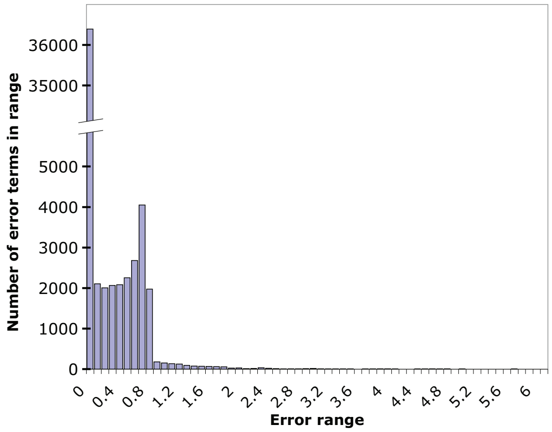

Figure 2: Histogram graph of the number of error ϵ_k terms with certain values. The vast majority are 0.

This is equivalent to:

$$\begin{aligned} -\mu d_1(k) + \gamma d_2(k) \leq&\quad -1 \quad \forall k | c_k = 1 \\ \mu d_1(k) - \gamma d_2(k) \leq&\quad 1 \quad \forall k | c_k = -1 \end{aligned}$$

Thus the training problem is to choose μ and γ so that

$$-\mu d_1(k) c_k + \gamma d_2(k) c_k \leq -c_k \tag{2}$$

is true for as many k as possible. The optimal solution would minimize the number of k for which Equation 2 is not true. To approximate this, we subtract an error term per frame and minimize the sum. The linear program is:

$$\min \sum_{k=1}^{n} \epsilon_k$$

subject to

$$\begin{aligned} -\mu d_1(1)c_1 + \gamma d_2(1)c_1 - \epsilon_1 &\leq -c_1 \\ -\mu d_1(2)c_2 + \gamma d_2(2)c_2 - \epsilon_2 &\leq -c_2 \\ \vdots \quad & \quad \vdots \\ -\mu d_1(n)c_n + \gamma d_2(n)c_n - \epsilon_n &\leq -c_n \\ \mu, \gamma, \epsilon_k &\geq 0 \end{aligned}$$

The variables in the linear program are μ, γ, and $\epsilon_k, 1 \leq k \leq n$. We normalize the $d_1(k)$ and $d_2(k)$ so that they are between 0 and 1 and run Simplex to find the settings for μ and γ that minimize the error. The classification of an unknown frame p is "signing" if $-\mu d_1(p) + \gamma d_2(p) \leq -1$ and "not signing" otherwise.

Though this is not an optimal protocol, in practice the error values ϵ_k are quite small. Figure 2 shows a histogram of ϵ values. The majority, over 66 percent of the total, are equal to 0, and the mean is 0.19.

Support vector machine. In the earlier work, the optimal method used features available "for free" from the H.264 encoder, plus skin information, and trained a support vector machine (SVM) for classification. The encoder features were a summary motion vector and the number of intra blocks. The skin features were the area, centroid, and bounding box of the three largest skin blobs, ideally corresponding to the face and hands. Previously, this was implemented on a standard PC using web camera videos. To implement skin feature extraction on the phone, we need to detect the skin rapidly, filter and threshold, and determine the three largest connected components and their area, center of gravity, and bounding box.

Since we are using relatively coarse features from only the three biggest skin patches, our skin detection algorithm does not have to be very precise. We detect the skin using a simple range query on the chrominance components. Yang et al. [32] found that when the color model was YUV, the skin is constrained to a small space in the chrominance dimension and is relatively robust to different lighting conditions. The differences in skin color due to race are also mainly captured in the luminance component. In our experiments, we assigned a pixel as skin if the U component was between 77 and 127 and the V component was between 133 and 173, and found it robust to differences in lighting and skin tone. We apply a 4 × 4 averaging filter on the binary skin map to eliminate small holes. We detect the connected components by using the classical two-pass labeling algorithm that employs a union-find data structure [23]. As we label the components, we keep track of their current area, center of gravity, and bounding box, so that after we've found the connected components, the task is finished; we do not need to iterate over the skin map again. This version of the feature extraction is cheap enough not to affect the encoding frame rate.

To improve the classification accuracy, we use the change in area, centroid, and bounding box of the three components, rather than their raw value. We also add the pixel differences from the person to whom we're talking. This is transmitted in packets whether or not the frame itself is sent. Since that data is so small, transmitting it does not affect the bit rate or encoding time.

Unfortunately, the float operations of the SVM cause an unacceptable delay in the encoding of the video. Although we could not use the algorithm for evaluation with users, we report on the results of testing it against the other two methods for classification accuracy.

Dynamic Skin Region-of-Interest Encoding

The encoder aims to produce the highest possible quality frame at a given bit rate (in our case, 30 kbps). The quality of each macroblock is determined by the quantizer step size, or *QP*. Lower QPs indicate higher quality but also require a higher bit rate.

We employ a simple skin detection technique on the uncompressed image. The image is captured in YUV format, where Y is the luminance component and U and V are the chrominance components. We examine the U and V values and determine if they are in the appropriate range for skin. We then check each 16 × 16 pixel macroblock and deem it skin if the majority of pixels in the block are skin. We change the quality of the skin by adjusting the QP value for the skin macroblocks. In our experiments, ROI 0 corresponds to no reduction in quantizer step size; ROI 6 corresponds to a 6 step reduction in size; and ROI 12 corresponds to a 12 step reduction in size. Forced to make the skin macroblocks of higher quality, the encoder must reduce the quality elsewhere in the frame to maintain the bit rate.

Figure 3 shows a comparison of ROI 0 (left) and ROI 12 (right). The face and hand are slightly clearer in the ROI 12 picture. However, in the ROI 0 picture there is a clear line between the shirt and the background around the shoulder area. In the ROI 12 picture, this line is smudged and blocky.

SYSTEM EVALUATION RESULTS

In this section, we describe the classification results and the power savings for the VFR. In the next section we evaluate the intelligibility of our system with users.

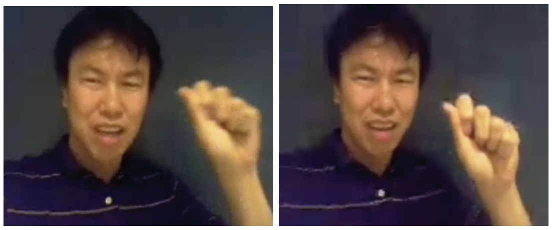

Figure 3: ROI 0 (left) and ROI 12 (right). Notice that the skin in the hand is clearer at ROI 12, but the background and shirt are far blurrier.

Classification Accuracy

To test the classification accuracy of our algorithms, we captured YUV video from the phone camera and hand labeled the frames as signing or not. We recorded four conversations with six different users, for a total of 8 videos. We evaluated the accuracy of the methods by dividing each video into four parts, training on three, and testing on the fourth. We report the average score. This is equivalent to a user spending time training the phone before first use. Since phones are personal devices, it would be natural to add this feature. To smooth out the results temporally, we apply a sliding window that takes the average vote over the window and classifies accordingly. We experimented with several different window sizes and found the three-frame window to be the best across all methods.

Figure 4 displays a comparison of the average classification accuracy of the different methods for each video. Over all the videos, SVM performed the best, with 77.8% of frames on average classified correctly. Baseline differencing also performed quite well, with 76.6% of frames on average classified correctly. Joint linear programming was a distant third, with 67.1% accuracy, even though using joint information had performed well in previous work. We believe that the

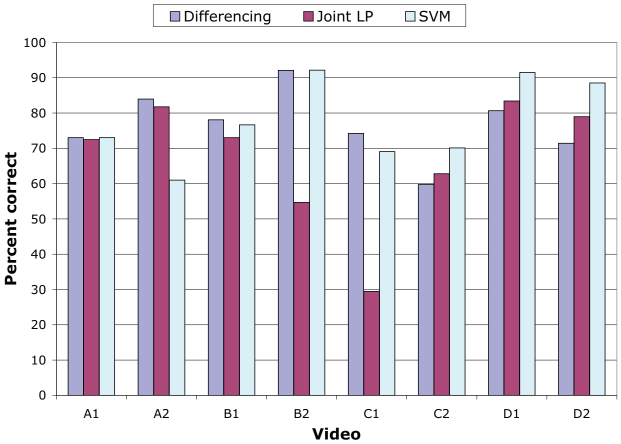

Figure 4: Classification accuracy on the eight videos with the three methods described in the previous section. On average, SVM was slightly better, but the results are quite similar. Differencing is used on the MobileASL phones.

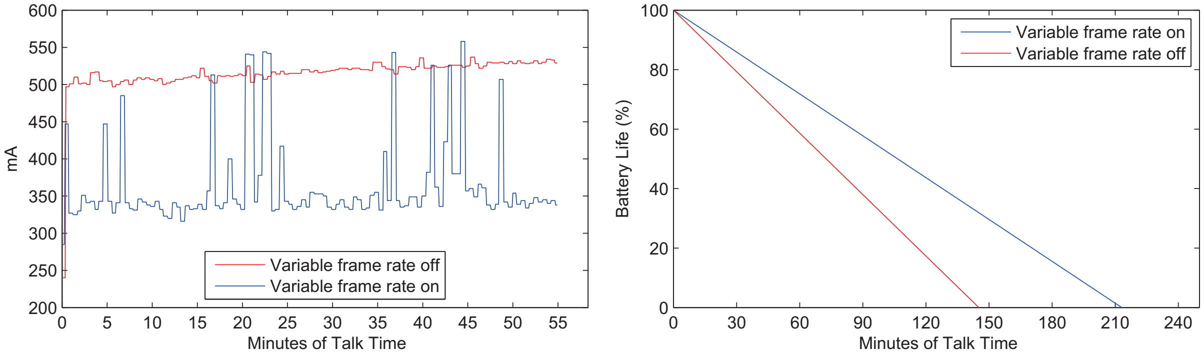

Figure 5: Power savings when encoding the video with the VFR turned on. (Left) Snap shot of the power draw in milliamps (mA). (Right) Battery drain. The VFR yields an additional 68 minutes of talk time.

smaller field-of-view on the phone videos compared to the earlier web cam videos affected the influence of the joint information, making it less relevant.

It is a welcome surprise that baseline differencing performs so well, because it is computationally cheap to implement on the phone. The SVM uses the differencing as a feature, but it is likely that it is overtraining to noisy data with the other features. Upon further investigation, we found that the sum of absolute differences is the most salient feature, followed by the motion vectors.

Power Savings

In order to quantify the power savings on the phone, we simulated a sign language conversation and monitored the power usage for an hour on two phones with the VFR on and with it off. The simulated conversation consisted of motion resulting in the higher frame rate every other minute, as though each person were signing for a minute, then listening for a minute, and so on. From observations of users on the phones, this seemed to be a reasonable scenario. Figure 5 shows a snap shot of the power draw when the phone utilizes a VFR versus when it does not. The power draw dips when the frame rate is lowered, due to the less processing power required to encode and transmit at 1 fps. Over an hour of talk time, the average power draw is 32% less with VFR on than with it off.

In terms of battery life, the power savings is dramatic. Testing two different phones over the course of an hour, the phone without the VFR lost 39% of battery life, versus 25% when the VFR was on. Regression analysis shows that the rate of loss over time for battery life on the phones is linear, with correlation coefficients of greater than 0.99. The average slope of the power drain on the battery every 5 seconds with the VFR off is -0.0574, versus -0.0391 with it on. This corresponds to 68 extra minutes of talk time, or a 47% power gain over the battery life of the phone (see Figure 5).

MOBILEASL USER STUDY

In this section, we evaluate the user experience of MobileASL in a laboratory setting. We tested VFR on and off together with dynamic skin ROI encoding. The settings for the ROI varied between 0, or no ROI; 6, or low ROI; and 12, or high ROI. The numbers correspond to encoder settings. The three different ROI levels and two different VFR settings result in six different possible combinations. The order of the settings was changed between conversations according to a Latin Square. Differencing with the τ parameter (Equation 1) trained on earlier videos was used to decide when to lower the frame rate. The signing frame rate of the phones was 7-8 fps. There was no perceptible delay. The architecture is shown in Figure 6.

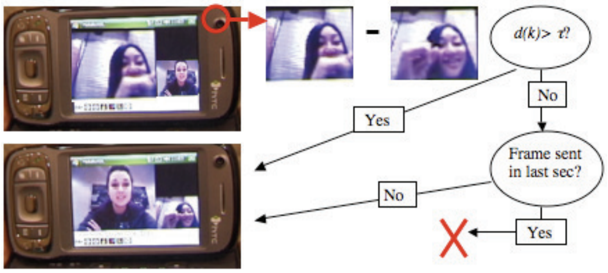

Figure 6: The architecture of the variable frame rate. Differences between frames are checked; if the user isn't signing, the frame is sent only to maintain 1 fps.

Given that our application is mobile phone communication, we expect a variety of different conversations to take place between people who may or may not already know each other. For example, a user might call an unfamiliar interpreter in order to reach his or her doctor. On the other hand, users will certainly call friends and family members. The conversations recorded represent this variety. There is always a tradeoff between repeatability of experiments and the realism of the setup; we erred on the side of realism.

We gathered both subjective and objective measures. The subjective measures were obtained via a survey. For the objective measures, we wanted to see how the conversations were affected by our changes to VFR and ROI. We videotaped each conversation and analyzed the recording after.

Participants

Altogether, 15 participants fluent in ASL (age: 24-59, mean = 42, 5 male) recruited from the Seattle area took part in the study. Eight of 15 participants preferred ASL for communication, four preferred English, and the remaining three chose both. Of the participants that chose English, three were interpreters with 7, 15, and 30 years experience, and one read lips but had known ASL for 21 years. Five participants wore glasses for near-sightedness.

Nine separate conversations were recorded. Three conversations were with a research staff member fluent in ASL, so statistics were only collected from one side of those conversations. Five of the conversations were between strangers and four were between people that knew each other well, including one husband/wife pair.

Apparatus

The participants sat on the same side of a table, separated by a screen. In the background was a black drape. The phones were on the table in front of them, and they were told to adjust their positioning and the phone location so that they were comfortable carrying on the conversation.

Procedure

The participants were told to talk about whatever came to mind, and that they would be interrupted after five minutes and the settings changed on the phone. After each five minute period (*subconversation*), they filled out a paper questionnaire. Each conversation was videotaped, and objective measures were calculated from the recording.

Subjective measures. The participants were asked to subjectively rate the quality of the video, measured by how hard or easy it was to understand. The survey questions were as follows:

- During the video, how often did you have to guess what the signer was saying (where 1 is never and 5 is all the time)?
- How difficult would you say it was to comprehend the video (where 1 is very easy and 5 is very difficult)?
- Changing the frame rate of the video can be distracting. How would you rate the annoyance level of the video (where 1 is not annoying at all and 5 is extremely annoying)?
- The video quality over a cell phone is not as good as video quality when communicating via the Internet (e.g., by using a web cam) or over a set top box. However, cell phones are convenient since they are mobile. Given the quality of conversation you just experienced, how often would you use the mobile phone for making video calls versus just using your regular version of communication (e.g., go home to use the Internet or set top box, or just text)?
- If video of this quality were available on the cell phone, would you use it?

The fourth question was poorly worded and often had to be explained verbally. We were trying to capture the trade-off in the greater convenience of the mobile phone than other methods versus its lower quality. After we explained the purpose of this question, participants understood it without confusion.

Objective measures. Our goal was to measure the comprehensibility of the conversation. A confusing conversation might contain a lot of requests for repetitions, called repair requests [28], and conversational breakdowns, where one person says, "I can't hear you," gives up, or similar. In sign language, there is an additional feature, which is finger spelling. Finger spelling is when someone spells out the name of something, and occurs mainly with proper names, titles, and technical words. However, some finger spelled words are lexicalized "loan signs," common words whose sign has become the stylized finger spelling (e.g., "bus," "back"). Since these act as regular signs, we do not count them in our finger spelling measure.

The objective measures were number of repair requests, average number of turns associated with repair requests, number of conversational breakdowns, and speed of finger spelling. These were all calculated from the videotaped user study sessions with the help of a fluent ASL speaker. In sign language conversation, a repair request may mean forming the sign for "again" or "what?," or finger spelling in unison with the conversation partner. For each repair request, we counted the number of turns until the concept was understood; this is the number of times the requester had to ask for repetition before moving on. Conversational breakdowns were calculated as the number of times the participant signed the equivalent of "I can't see you" (e.g. "frozen," "blurry," "choppy"). If repair requests were made but never resolved, we counted it as a conversational breakdown. Finally, we measured the time it took to sign each finger spelled word and divided by the number of characters in that word, resulting in characters per second.

User Study Results

The results of the study were statistically significant for only one of the subjective measures, guessing. VFR affected all of the objective measures except for finger spelling speed but ROI did not.

Likert scale subjective measures. Table 1 contains the χ^2 test and significance values for the five questions. Only the first question in the questionnaire yielded statistically significant results. The interaction results were all non-significant, indicating levels of ROI and VFR did not disproportionately affect one another.

The ROI had a significant effect on participants' Likert responses for how often they had to guess, with 1=not at all, 5=all the time. A Wilcoxon signed-rank test shows that ROI 0 and ROI 6 were not significantly different $(z = 0.50, ns)$, but that ROI 0 and ROI 12 were different $(z = 35.00, p < .01)$ and ROI 6 and ROI 12 were also different $(z = 35.50, p < .05)$. Thus, perceptions of guessing frequency decreased when ROI coding reached 12 from 0 and 6. The VFR increased perceptions of guessing frequency. The means for ROI 0, 6, and 12 were 1.90, 1.88, and 1.42, respectively. The mean for VFR off was 1.60 and for VFR on was 1.87.

The ROI and VFR did not cause a detectable difference in participants' Likert responses for how annoyed they were at the level of frame rate, how often they would prefer the phone to some other means of communication, or their potential future use of the technology. The overall means

Question	ROI		VFR		Interaction	
	$\chi^2(2, N = 90)$	p	$\chi^2(1, N = 90)$	p	$\chi^2(2, N = 90)$	p
Guesses	11.11	$< .01^{**}$	4.44	$< .05^{*}$	0.78	.68
Comprehension	5.33	.07	2.87	.09	2.75	.25
Annoyance	3.07	.22	0.79	.37	2.26	.32
Phone vs. Other	0.12	.94	1.10	.29	0.18	.91
Would use	0.42	.81	0.22	.64	1.02	.60

Table 1: Statistical analysis for the subjective measures. Statistical significance: ** = $p < 0.01$, * = $p < 0.05$.

Question	ROI		VFR		Interaction	
	$\chi^2(2, N = 90)$	p	$\chi^2(1, N = 90)$	p	$\chi^2(2, N = 90)$	p
Repair requests	2.66	.26	5.37	$< .05^{*}$	1.99	.37
Number of turns	0.94	.62	4.01	$< .05^{*}$	0.96	.62
Breakdowns	3.38	.18	7.82	$< .01^{**}$	1.51	.47
	$F(2, 28)$	p	$F(1, 14)$	p	$F(2, 28)$	p
Finger spelling speed	0.42	.66	0.19	.67	0.21	.81

Table 2: Statistical analysis for the objective measures. Statistical significance: ** = $p < 0.01$, * = $p < 0.05$.

for preference of the phone and potential future use were 2.98 and 2.47, respectively, where 1 means the participant thinks they would definitely use the phone and 5 means the participant thinks they would definitely not use the phone.

Objective measures. Table 2 contains the statistical results for the objective measures. Repair requests, number of turns before a concept was understood, and conversational breakdowns were all affected significantly by VFR, but not by the ROI. Speed of finger spelling was affected by neither, and the interaction between ROI and VFR was not statistically significant. The number of repair requests was highly skewed and according to a significant Shapiro-Wilk test ($W = 0.74$, $p < .0001$), not amenable to ANOVA. This was also true of the number of repetitions that transpired before the concept was understood ($W = 0.76, p < .0001$) and the number of conversational breakdowns ($W = 0.44, p < .0001$). Typical corrective transformations were not an option, so we continued to employ ordinal logistic regression as we had for our Likert data, which showed an appropriate fit for all three measures. For finger spelling speed, a non-significant Shapiro-Wilk test ($W = 0.98, p = .12$) confirms these data are suited to analysis with a repeated measures ANOVA. Finger spelling speed was not affected significantly by ROI or VFR, and the mean finger spelling speed for all conditions was 3.28 characters per second.

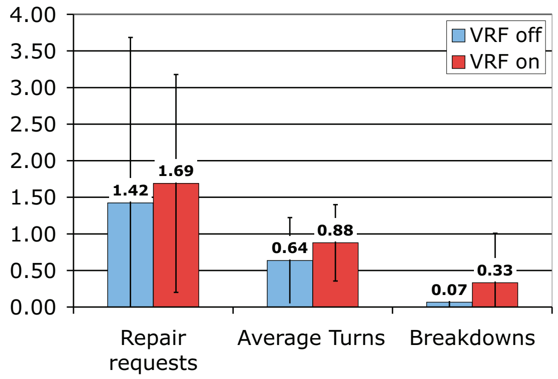

Figure 7: The number of repair requests, the average number of turns to correct a repair request, and the conversational breakdowns.

The means and standard deviations for the significant objective measures are in Figure 7. VFR negatively affects the number of repeats, the number of repetitions, and the number of conversational breakdowns.

Participant comments. Nearly all of the participants asked when the technology would be available for their use. They expressed disappointment that the software was not ready for widespread distribution.

Several participants commented on the awkward angle of the camera when the phone is on the table. They separately suggested creating a stand so that the phone could be at the same level as the face. Two of the participants disliked the eye strain caused by looking at the small screen. Participants 8 and 9 were affected by an overly bright room that made the LCD very difficult to see. Participant 9 wore glasses for near sightedness. They commented that they would only use the phone for emergencies or very short conversations.

Two of the participants who were interpreters separately commented on the speed of finger spelling. They noted that they were finger spelling at a pace somewhat slower than usual and said this reminded them of video-relay interpreting. Video-relay interpreting occurs over a much higher bandwidth connection than the mobile phone, but it sometimes has connection problems and can induce similar human behavior.

DISCUSSION

Our participants felt that they had to guess less frequently at higher ROI levels. ROI otherwise had no statistically significant effect on the participants. Recall that a high ROI encodes the skin at a higher quality at the expense of the rest of the frame, meaning there is no extra cost to the encoder in terms of bandwidth. Since our algorithm is a simple range-detection query, there is no extra burden on the processor. Thus, using a high ROI is a good way to save system resources and still increase intelligibility.

The results on VFR are more mixed. We expect VFR to lead to some degradation in quality, since we are leaving out a lot of information in order to save resources. Indeed, participants felt they had to guess more frequently. Moreover, their conversations were also objectively affected. They made more repair requests, took more turns to correct those requests, and had more conversational breakdowns when VFR was on. However, an examination of means shows that overall, they were not making many repair requests or experiencing many conversational breakdowns. Breakdowns only occurred once every third conversation on average.

The results on three of the subjective measures were encouraging. The VFR did not appear to affect participants' feelings that they would use the phone instead of other means of communication or that they would adopt the phone in general. It also did not affect their irritation with the frame rate changes. Because VFR saves considerable system resources, we expect it to affect conversations; it is encouraging that this does not mean users feel that they are less likely to adopt the technology. In a laboratory setting, it is difficult to capture the advantage of a long-lasting battery. Our future field study will allow participants to use the phones over a longer period of time, so we can better evaluate the trade-off in battery life versus VFR.

The results on finger spelling were surprising. Given the other objective measure results, we expected finger spelling to be measurably slower with VFR turned on, but we saw no statistically significant difference. It may be that the participants spelled more slowly overall and not just during the VFR subconversations. However, when analyzing videos it seemed to us that conversational breakdowns occurred most often when one participant was finger spelling. We suspect this is because the differencing method would incorrectly label the frames and lower the frame rate, occasionally resulting in a "frozen" image.

FUTURE WORK

Although we solved many hard technical problems, several other technical challenges remain. We would like to further investigate finger spelling. Using our method developed for the VFR, we want to automatically recognize finger spelling so that we do not lower the frame rate during these periods of the video. It would also be interesting to know how using the mobile phone affects finger spelling compared to other methods of video communication, such as a video relay service. In general, we would like to better model sign language as opposed to just motion, as we do now. We cannot currently distinguish between extraneous motion, such as someone drinking coffee, and the purposeful motion of signing.

There are also several different ways we might improve our classification. We could use a different machine learning algorithm like boosting; choose different features, such as histograms of motion vectors, to send to our classifier; and try to speed up the SVM classifier through quantization. Face detection algorithms are quite fast, and are implemented in digital cameras, so it might be possible to make the dynamic ROI track the face. Adding more training data would improve classification.

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.

CONCLUSION

In this work, we describe our system for real time video communication over mobile phones. We create techniques that save system resources including processor workload and battery life by focusing on the important parts of the video. We implement our methods on an off-the-shelf mobile phone and evaluate our techniques in a user study in which participants carry on unconstrained conversation over the phones in a laboratory setting.

The most common question asked by our participants was "when will this be available?" When recruiting for our study, we received interested queries from all over the United States. We are encouraged by the results of this work as it furthers our ultimate goal: to provide Deaf people full access to today's mobile telecommunication network.

ACKNOWLEDGEMENTS

The authors thank Howie Seago, Ginevra Deianni, Jessica DeWitt, and Gina Yim for their help in running the user study and creating and analyzing videos. This work was supported in part by the National Science Foundation under grants IIS-0811884 and CCF-0514353. Any opinions, findings, conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect those of the National Science Foundation.

REFERENCES

1. 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.

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

3. D. Bavelier, A. Tomann, C. Hutton, T. Mitchell, D. Corina, G. Liu, and H. Neville. Visual attention to the periphery is enhanced in congenitally deaf individuals. *The Journal of Neuroscience*, 20(RC93):1–6, 2000.

4. N. Cherniavsky, R. E. Ladner, and E. A. Riskin. 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, Sept 2008.

5. E. Clarkson, J. Clawson, K. Lyons, and T. Starner. An empirical study of typing rates on mini-QWERTY keyboards. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1288–1291, 2005.

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

7. L. Garber. Technology news: Will 3G really be the next big wireless technology? *Computer*, 35(1):26–32, January 2002.

8. GSMA. General packet radio service. `http://www.gsmworld.com/technology/gprs/class.shtml`, 2006.

9. 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.

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

11. R. Hsing and T. P. Sosnowski. Deaf phone: sign language telephone. In *SPIE volume 575: Applications of digital image processing VIII*, pages 56–61, 1985.

12. International Telecommunication Union. International Mobile Telecommunications-2000 (IMT-2000), 2000. http://www.itu.int/home/imt.html.

13. International Telecommunication Union. Trends in Telecommunication Reform 2007: The Road to NGN, Sept 2007.

14. 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.

15. B. F. Johnson and J. K. Caird. 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, New York, NY, USA, 1996. ACM Press.

16. L. Merritt and R. Vanam. Improved rate control and motion estimation for H.264 encoder. In *Proceedings of ICIP*, volume 5, pages 309–312, 2007.

17. Joint Model. JM ver. 10.2. http://iphome.hhi.de/suehring/tml/index.htm.

18. L. Muir and I. Richardson. Perception of sign language and its application to visual communications for deaf people. *Journal of Deaf Studies and Deaf Education*, 10(4):390–401, 2005.

19. S.C.W. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6), June 2005.

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

21. D. E. Pearson. Visual communication system for the deaf. *IEEE Transactions on Communication*, 29:1986–1992, December 1981.

22. C. M. Reed, L. A. Delhorne, N. I. Durlach, and S. D. Fischer. A study of the tactual and visual reception of fingerspelling. *Journal of Speech and Hearing Research*, 33:786–797, December 1990.

23. A. Rosenfeld and J. Pfaltz. Sequential operations in digital picture processing. *Journal of the ACM*, 13(4):471–494, 1966.

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

25. 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.

26. G. Sperling, M. Landy, Y. Cohen, and M. Pavel. 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, 1986. Academic Press Professional, Inc.

27. W. C. Stokoe. *Sign Language Structure: An Outline of the Visual Communication System of the American Deaf.* Studies in Linguistics: Occasional Papers 8. Linstok Press, Silver Spring, MD, 1960. Revised 1978.

28. D. R. Traum and E. A. Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8:575–599, 1992.

29. M. A. Viredaz and D. A. Wallach. Power evaluation of a handheld computer. *IEEE Micro*, 23(1):66–74, 2003.

30. T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Techn*, 13(7):560–576, 2003.

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

32. J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaptation. In *Proceedings of the Third Asian Conference on Computer Vision-Volume II*, pages 687 – 694. Springer-Verlag, 1998.