Variable Frame Rate for Low Power Mobile Sign Language Communication

Neva Cherniavsky Anna C. Cavender Richard E. Ladner Eve A. Riskin†

Department of Computer Science and Engineering, Box 352350,
† Department of Electrical Engineering, Box 352500,
University of Washington, Seattle, WA 98195.
Email: {nchernia, cavender, ladner}@cs.washington.edu
{riskin}@ee.washington.edu

ABSTRACT

The MobileASL project aims to increase accessibility by enabling Deaf people to communicate over video cell phones in their native language, American Sign Language (ASL). Real-time video over cell phones can be a computationally intensive task that quickly drains the battery, rendering the cell phone useless. Properties of conversational sign language allow us to save power and bits: namely, lower frame rates are possible when one person is not signing due to turn-taking, and signing can potentially employ a lower frame rate than fingerspelling. We conduct a user study with native signers to examine the intelligibility of varying the frame rate based on activity in the video. We then describe several methods for automatically determining the activity of signing or not signing from the video stream in real-time. Our results show that varying the frame rate during turn-taking is a good way to save power without sacrificing intelligibility, and that automatic activity analysis is feasible.

Categories and Subject Descriptors

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

General Terms

Human Factors

Keywords

Low Power, Activity Analysis, Sign Language, Deaf Community, Mobile Telephone Use

1. INTRODUCTION

Mobile phones with the ability to display, capture, and transmit video are becoming more widespread in the marketplace. These phones will soon enable better access to the mobile cell phone network for people within the signing Deaf Community. While many of the approximately one million Deaf people in the U.S. [15] are already using internet based video phones, there is currently no equivalent form of communication over the mobile phone network in the U.S. This is partly due to bandwidth constraints [12] and partly due to limited processing power of phones when required to compress video at such low bit rates. As mobile phone networks improve (for example, 3G technology is available in several countries such as Sweden and Japan and some major cities in the U.S. [1]) and video compression techniques advance, the challenge will shift from minimizing the bit rate to minimizing the processor load. As part of the MobileASL project [3, 6], we are developing video encoding techniques that reduce both computation and bandwidth without significantly harming sign language intelligibility.

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 can destroy its usefulness if a conversation cannot last 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 [20]. For a sign language conversation, not only do we want to play video, but also we want to capture, encode, transmit, receive and decode video all at once and all in real-time. Needless to say, we can expect battery life to be even more quickly depleted.

One way to save battery life is to encode videos at a lower frame rate (i.e. encoding fewer frames per second). Decreasing the frame rate reduces the average number of processor cycles needed (see Figure 1) and reducing cycles helps save power. Previous studies have shown that when playing video, 30% of the power consumption is due to the processor [20]. Not only does encoding fewer frames save power, *sending* fewer frames saves power. Several studies have shown that the transmit mode consumes more power than the receive mode [5, 10]. Sending fewer frames also reduces the total bandwidth consumed by the user and helps reduce the load on the network. Depending on the pricing model, the user could also benefit; for example, if the company bases its fees on the amount of data transmitted, a lower frame rate would result in a cheaper bill for the user.

Our goal is to enable real-time mobile sign language con-

ASSETS'07, October 15–17, 2007, Tempe, Arizona, USA.

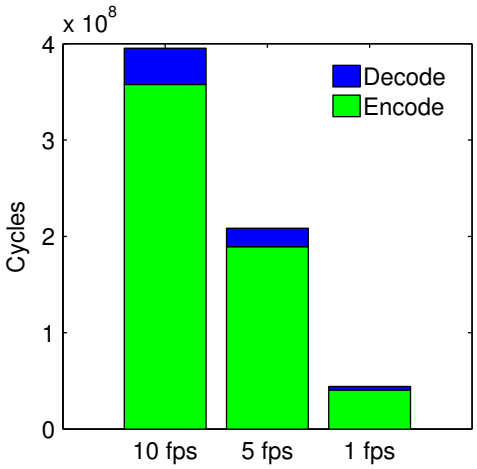

Figure 1: Average processor cycles per second for a video encoded at 10 frames per second, 5 frames per second, and 1 frame per second.

versations and part of that goal will be minimizing the frame rate of video transmitted. However, we do not want to send a video at such a low frame rate that it becomes unintelligible. Previous work has shown that frame rates as low as 6 frames per second can be intelligible for signing, but higher frame rates are needed for fingerspelling [11, 19, 13]. In our work, we leverage the natural structure of two-sided conversations as well as linguistic aspects of sign language, such as fingerspelling, that may require more or less temporal information. Because conversation involves turn-taking (times when one person is signing while the other is not), we can save power as well as bit rate by lowering the frame rate during times of not signing, or "just listening" (see Figure 2). We can also try to increase intelligibility by increasing the frame rate during fingerspelling.

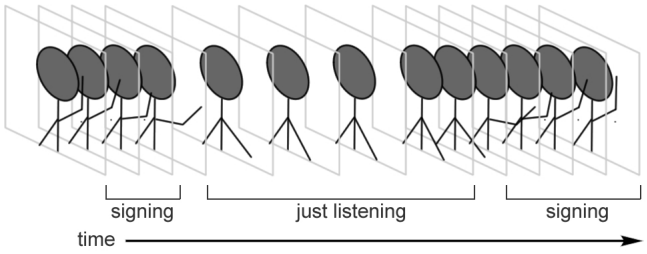

Figure 2: From left to right: a sufficient video frame rate is chosen when the signer is signing, the frame rate decreases when the signer is not signing (or just listening), and increases again when the signer begins signing.

In this work, we examine the feasibility of adjusting the frame rate for different activity in the video. We first describe a user study in which Deaf participants evaluate videos with differing frame rates. We then describe methods for determining the signer's activity from only the information available in the video in order to automatically find appropriate times to adjust the frame rate.

2. RELATED WORK

The intelligibility of varying frame rates depending on video content has not, to our knowledge, been examined. Conversational sign language is similar to spoken language in that multiple people may "hold the floor" at once [7]. Furthermore, ASL contains back-channel feedback [8], in which the listener indicates understanding, similar to a hearing person saying "uh-huh." Since users of MobileASL may be "signing over one another," we want to know if reducing the frame rate when one user is not signing negatively affects intelligibility.

A related topic is sign language recognition, in which researchers try to translate sign language into English text. Several good surveys detail the state-of-the-art [16, 14]. However, the goal of our project does not involve translation or interpretation. Instead, we aim to increase accessibility by enabling Deaf people to communicate over cell phones. The domain of mobile communication restricts potential solutions to those that can utilize only video and that are computationally simplistic enough to run in real-time on limited mobile phone processors.

Johnson and Caird investigated the effects of frame rate on sign language instruction [13]. They found that 1 and 5 frames per second (fps) were sufficient for novices to learn from ten ASL videos, each containing one sign. The effect of frame rate on intelligibility of isolated ASL signs has also been studied by Sperling et al. who found insignificant differences in intelligibility from 30 to 15 fps, slight reduction in intelligibility from 15 to 10 fps, and considerable reduction from 10 to 5 fps [19]. Foulds similarly found 6 fps can accurately represent ASL and fingerspelling for individual signs when smoothly interpolated to 30 fps [11]. Since our videos contain more "conversationally-paced" signing with many rapidly-produced signs, and our users could be considered experts in sign language, 5 fps is likely a lower bound for sufficient comprehension.

Automatic activity analysis of video is an active topic of research in the computer vision community. While conversational sign language video is not widely studied, there are several related problems that have received attention. Shot change detection [17] determines when a scene changes in a video, so that it can be parsed automatically and key frames extracted. There is usually no need for real-time analysis in shot change detection, so most algorithms analyze the entire video at once. Furthermore, there are usually substantial differences between scenes, while in our videos there are only minor differences between the signing and not signing portions. Our baseline differencing method is a common starting point for shot change detection. Another related area is human motion analysis [21]. Usually the goal of motion analysis is to track or recognize people or activities from video. Often the computer vision techniques are not real-time, and require processing power far beyond the scope of a mobile phone.

3. STUDY DESIGN

To better understand intelligibility effects of altering the frame rate of sign language videos based on language content, we conducted a user study with members of the Deaf Community. The purpose of the study was to investigate the effects of (a) lowering the frame rate when the signer is not signing (or "just listening") and (b) increasing the

frame rate when the signer is fingerspelling. The hope was that study results would motivate the implementation of our proposed automatic techniques for determining conversationally appropriate times for adjusting frame rates in real time with real users.

The videos used in our study were recordings of conversations between two local Deaf women at their own natural signing pace. During the recording, the two women alternated standing in front of and behind the camera so that only one person is visible in a given video. The resulting videos contain a mixture of both signing and not signing (or "just listening") so that the viewer is only seeing one side of the conversation. The effect of variable frame rates was achieved through a "Wizard of Oz" method by first manually labeling video segments as signing, not signing, and fingerspelling and then varying the frame rate during those segments.

Even though the frame rate varied during the videos, the bits allocated to each frame were held constant so that the perceived quality of the videos would remain as consistent as possible across different encoding techniques. This means that the amount of data transmitted would decrease with decreased frame rate and increase for increased frame rate. The maximum bit rate was 50 kbps.

We wanted each participant to be able to view and evaluate each of the 10 encoding techniques described below without watching the same video twice and so we created 10 different videos, each a different part of the conversations. The videos varied in length from 0:34 minutes to 2:05 minutes (mean = 1:13) and all were recorded with the same location, lighting conditions, and background. The x264 codec [2], an open source implementation of the H.264 (MPEG-4 part 10) standard [18], was used to compress the videos.

Both videos and interactive questionnaires were shown on a Sprint PPC 6700, PDA-style video phone with a 320 × 240 pixel resolution (2.8" × 2.1") screen.

3.1 Signing vs. Not Signing

We studied four different frame rate combinations for videos containing periods of signing and periods of not signing. Previous studies indicate that 10 frames per second (fps) is adequate for sign language intelligibility, so we chose 10 fps as the frame rate for the signing portion of each video. For the non-signing portion, we studied 10, 5, 1, and 0 fps. The 0 fps means that one frame was shown for the entire duration of the non-signing segment regardless of how many seconds it lasted (a freeze-frame effect). Figure 3 shows the average cycles per second required to encode video using these four techniques and the savings gained from reducing the frame rate during times of not signing. A similar bit rate savings was observed; on average, there was a 13% savings in bit rate from 10-10 to 10-5, a 25% savings from 10-10 to 10-1, and a 27% savings from 10-10 to 10-0.

3.2 Signing vs. Fingerspelling

We studied six different frame rate combinations for videos containing both signing and fingerspelling. Even though our previous studies indicate that 10 fps is adequate for sign language intelligibility, it is not clear that that frame rate will be adequate for the fingerspelling portions of the conversation. During fingerspelling, many letters are quickly produced on the hand(s) of the signer and if fewer frames are

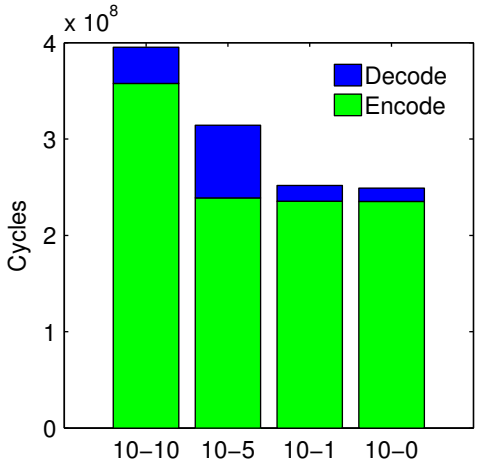

Figure 3: Average processor cycles per second for the four different variable frame rates. The first number is the frame rate during the signing period and the second number is the frame rate during the not signing period.

shown per second, critical letters may be lost. We wanted to study a range of frame rate increases in order to study both the effect of frame rate and *change* in frame rate on intelligibility. Thus, we studied 5, 10, and 15 frames per second for both the signing and fingerspelling portions of the videos resulting in six different combinations for signing and fingerspelling: 5 and 5, 5 and 10, 5 and 15, 10 and 10, 10 and 15, and 15 and 15.

3.3 Study Procedure

Six adult, female members of the Deaf Community between the ages of 24 and 38 participated in the study. All six were Deaf and had life-long experience with ASL; all but one (who used Signed Exact English in grade school and learned ASL at age 12) began learning ASL at age 3 or younger. All participants were shown one practice video to serve as a point of reference for the upcoming videos and to introduce users to the format of the study. They then watched 10 videos: one for each of the encoding techniques described above.

Following each video, participants answered a five- or six-question, multiple choice survey about his or her impressions of the video (see Figure 4). The first question asked about the content of the video, such as "Q0: What kind of food is served at the dorm?" For the Signing vs. Fingerspelling videos, the next question asked "Q1: Did you see all the finger-spelled letters or did you use context from the rest of the sentence to understand the word?" The next four questions asked:

Q2: "During the video, how often did you understand what the signer was saying?"

Q3: "How easy or how difficult was it to understand the video?"

Q4: "Changing the frame rate of the video can be distracting. How would you rate the annoyance level of the video?"

Q5: "If video of this quality were available on the cell phone, would you use it?"

The viewing order of the different videos and different encoding techniques for each part of the study (four for Signing vs. Not Signing and six for Signing vs. Fingerspelling) was determined by a Latin squares design to avoid effects of learning, fatigue, and/or variance of signing or signer on the participant ratings. Post hoc analysis of the results found no significant differences between the ratings of any of the 10 conversational videos. This means we can safely assume that the intelligibility results that follow are due to varied compression techniques rather than other potentially confounding factors (e.g. different signers, difficulty of signs, lighting or clothing issues that might have made some videos more or less intelligible than others).

4. RESULTS

For the variable frame rates studied here, we did not vary the quality of the frames and so the level of distortion was constant across test sets. Thus, one would expect to see higher ratings for higher frame rates. Our hope was that the ratings would not be statistically significant meaning that our frame rate conservation techniques do not significantly harm intelligibility.

4.1 Signing vs. Not Signing

For all of the frame rate values studied for non-signing segments of videos, survey responses did not yield a statistically significant effect on frame rate. This means that we did not detect a significant preference for any of the four reduced frame rate encoding techniques studied here, even in the case of 0 fps (the freeze frame effect of having one frame for the entire non-signing segment). Numeric and graphical results can be seen in Table 1 and Figure 4. This result may indicate that we can obtain savings by reducing the frame rate during times of not signing without significantly affecting intelligibility.

Many participants anecdotally felt that the lack of feedback for the 0 fps condition seemed conversationally unnatural; they mentioned being uncertain about whether the video froze, the connection was lost, or their end of the conversation was not received. For these reasons, it may be best to choose 1 or 5 fps, rather than 0 fps, so that some of feedback that would occur in a face to face conversation is still available (such as head nods and expressions of misunderstanding or needed clarification).

4.2 Signing vs. Fingerspelling

For the six frame rate values studied during fingerspelling segments, we did find a significant effect of frame rate on participant preference (see Table 2). As expected, participants preferred the encodings with the highest frame rates (15 fps for both the signing and fingerspelling segments), but only slight differences were observed for videos encoded at 10 and 15 fps for fingerspelling when 10 fps was used for signing. Observe that in Figure 4, there is a large drop in ratings for videos with 5 fps for the signing parts of the videos. In fact, participants indicated that they understood only slightly more than half of what was said in the videos encoded with 5 fps for the signing parts (Q2). The frame rate during signing most strongly affected intelligibility, whereas the frame rate during fingerspelling seemed to have a smaller effect on the ratings.

This result is confirmed by the anecdotal responses of study participants. Many felt that the increased frame rate during fingerspelling was nice, but not necessary. In fact many felt that if the higher frame rate were available, they would prefer that during the entire conversation, not just during fingerspelling. We did not see these types of responses in the Signing vs. Not Signing part of the study, and this may indicate that 5 fps is just too low for comfortable sign language conversation. Participants understood the need for bit rate and frame rate cutbacks, yet suggested the frame rate be higher than 5 fps if possible.

These results indicate that frame rate (and thus bit rate) savings are possible by reducing the frame rate when times of not signing (or "just listening") are detected. While increased frame rate during fingerspelling did not have negative effects on intelligibility, it did not seem to have positive effects either. In this case, videos with increased frame rate during fingerspelling were more positively rated, but the more critical factor was the frame rate of the signing itself. Increasing the frame rate for fingerspelling would only be beneficial if the base frame rate were sufficiently high, such as an increase from 10 fps to 15 fps. However, we note that the type of fingerspelling in the videos was heavily context-based; that is, the words were mostly isolated commonly fingerspelled words, or place names that were familiar to the participants. This result may not hold for unfamiliar names or technical terms, for which understanding each individual letter would be more important.

In order for these savings to be realized during real time sign language conversations, a system for automatically detecting the time segments of "just listening" is needed. The following section describes a potential solution.

5. VIDEO PROCESSING

We would like to automatically detect from our video stream when the user is signing, not signing, and fingerspelling, so we can lower or raise the frame rate accordingly. In this paper we tackle the question of automatically recognizing when the user is signing versus not signing, leaving the harder problem of fingerspelling detection to future work. For the purposes of frame rate variation, we can only use the information available to us from the video stream. We also must be able to determine the class of activity in real time.

We used the same four conversational videos from the user study. In each video, the same signer "Gina" is filmed by a stationary camera, and she is signing roughly half of the time. We are thus using an easy case as our initial attempt, but if our methods do not work well here, they will not work well on more realistic videos. We used three different techniques to classify each video into signing and not signing portions. In all the methods, we train on three of the videos and test on the fourth. We present all results as comparisons to the ground truth "Wizard of Oz" labeling.

5.1 Differencing

A baseline method is to examine the pixel differences between successive frames in the video. If frames are very different from one to the next, that indicates a lot of activity and thus that the user might be signing. On the other hand, if the frames are very similar, there is not a lot of

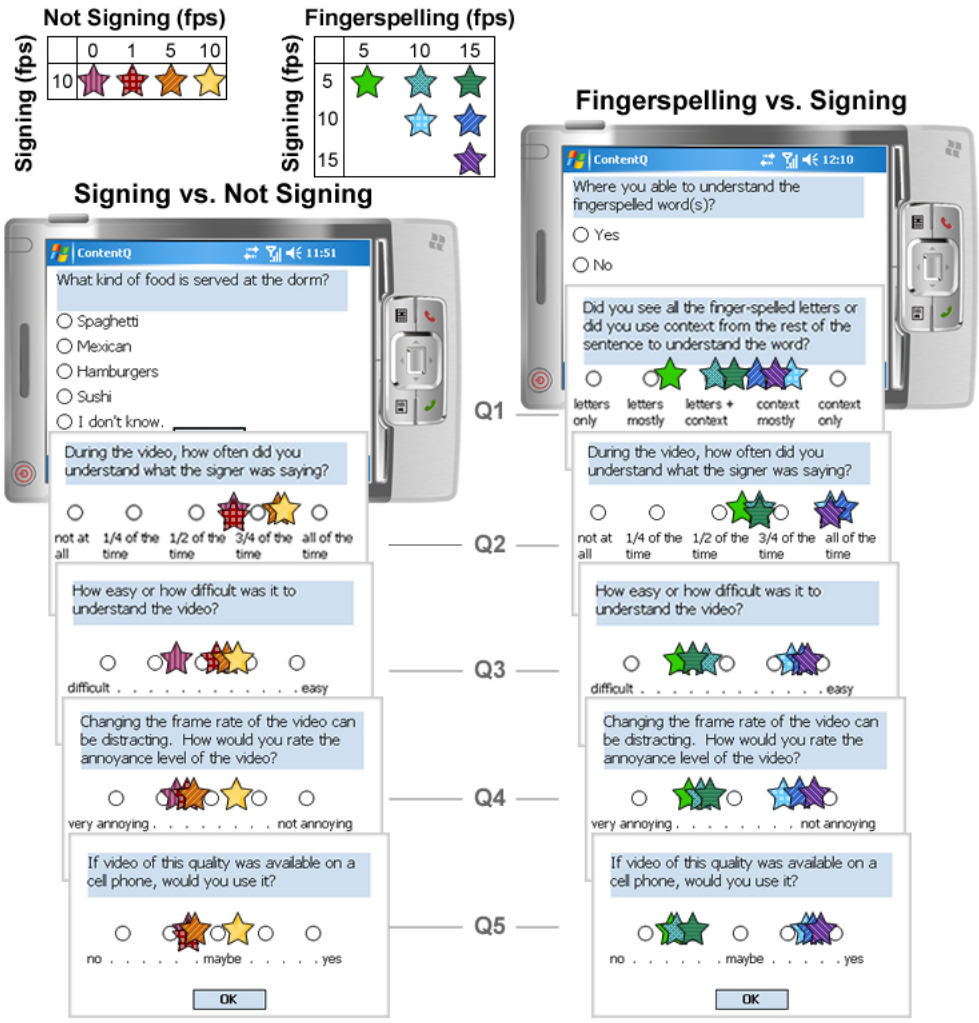

Figure 4: Average ratings on survey questions for variable frame rate encodings (stars).

Signing v Not Signing (fps)	10 v 0	10 v 1	10 v 5	10 v 10	Sig ($F_{3,15}$)
Q2 (0 not at all...1 all the time)	0.71 SD=1.88	0.71 SD=0.10	0.79 SD=0.19	0.83 SD=0.20	$1.00, n.s.$
Q3 (1 difficult...5 easy)	2.50 SD=1.64	3.17 SD=0.98	3.50 SD=1.05	3.83 SD=1.17	$1.99, n.s.$
Q4 (1 very...5 not annoying)	2.17 SD=1.33	2.50 SD=1.05	2.83 SD=1.33	3.67 SD=1.51	$1.98, n.s.$
Q5 (1 no...5 yes)	2.33 SD=1.75	2.33 SD=1.37	2.50 SD=1.52	3.33 SD=1.37	$1.03, n.s.$

Table 1: Average participant ratings for videos with reduced frame rates during non-signing segments.

motion so the user is probably not signing. As each frame is processed, it is subtracted from the previous frame, and if the differences in pixel values are above a certain threshold, the frame is classified as a signing frame. This method is sensitive to extraneous motion and is thus not a good general purpose solution, but it gives a good baseline from which to improve.

Formally, for each frame k in the video, we obtain the luminance component of each pixel location (i, j). We subtract from it the luminance component of the previous frame at the same pixel location. If the sum of absolute differences is above the threshold τ, we classify the frame as signing. Let $f(k)$ be the classification of the frame and $I_k(i, j)$ be the luminance component of pixel (i, j) at frame k. Call the difference between frame k and frame $k-1$ $d(k)$, and let $d(1) = 0$. Then:

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

$$f(k) = \begin{cases} 1 & \text{if } d(k) > \tau \\ -1 & \text{otherwise} \end{cases} \quad (2)$$