# Achieving Effective Floor Control with a Low-Bandwidth Gesture-Sensitive Videoconferencing System

Milton Chen
Computer Systems Laboratory
Stanford University
miltchen@stanford.edu

## ABSTRACT
Multiparty videoconferencing with even a small number of people is often infeasible due to the high network bandwidth required. Bandwidth can be significantly reduced if most of the advantages of using full-motion video can be achieved with low-frame-rate video; unfortunately, the impact of low-frame-rate video on communication is relatively unexplored. We implemented a multiparty videoconferencing system that supports full-motion video, low-frame-rate video where the video is updated only once every few seconds, and a hybrid scheme where full-motion video is transmitted when the system detects that a user is making a gesture and low-frame-rate video is transmitted at all other times. We studied people using our system for small-group discussions and found that low-frame-rate video limited people's ability to request to speak or judge when to stop speaking. The hybrid scheme, conversely, was as effective as full-motion video for floor control, resulting in a similar number of speaker changes, while using only ten percent of the bandwidth.

## Keywords
Multiparty videoconferencing, floor control, frame rate

## 1. INTRODUCTION
The most popular approach to synchronous distance learning today is to broadcast the instructor and the visual aids through a television network or the Internet [21]. To promote classroom interaction, students can talk to the instructor through a telephone, but the instructor and other students cannot see the remote students. While this approach has been used to educate thousands of remote students at Stanford University [28], a survey of 41 faculty members found that instructors dislike teaching students whom they cannot see [4]. Classroom observations further reveal that there is essentially no interaction with the remote students [5].

A visual feedback channel from the remote students to the instructor may promote greater classroom interaction [22]. The feedback channel can be used for both awareness and floor control. Awareness of the students' facial expressions, gestures, and postures allows an instructor to adapt the teaching to the students' current interest and understanding. Floor control, typically expressed through hand raising, allows students to indicate a desire to speak. The feedback channel can be based on (1) the text medium, such as Instant Messenger or chat [14][17], (2) the graphics medium, such as iconic representation of communication events [11] [14], or (3) the video medium [5][14].

Text and graphics feedback channels require very little network bandwidth, but students must perform explicit actions to communicate. For example, they may have to press a key to trigger a hand icon to indicate the desire to speak or click on an emoticon to express a puzzled look. Usage studies suggest that ephemeral feedback such as a fleeting smile or feedback that has a rigid timing requirement such as laughter after a joke may not be transmitted if explicit action is required [11]. People are also reluctant to explicitly express negative attitudes toward another [22]. Instructors are thus unlikely to see emoticons indicating that students are bored. An additional problem is that text and iconic channels do not transmit the appearance of the participants, a cue that is important when people interact with strangers [22], as is the case in many class settings.

A video feedback channel does not require participants to make all communicative actions explicit and conveys the appearance of the participants; however, the high network bandwidth required to stream full-motion video limits its deployment.

The goal of our research is to explore whether it is possible to achieve most of the benefits of full-motion video at significantly lower frame rates for remote

classrooms. Our hypothesis is that the visual cues necessary for classroom interaction do not need to be updated at the same rate. For example, while full-motion video is necessary for seeing a fleeting facial expression, low-frame-rate video may suffice for seeing posture changes. While floor control signals may require immediate transmission, delayed delivery of awareness cues may still have value.

To test our hypothesis, we implemented a multiparty videoconferencing system that supports full-motion video, low-frame-rate video where the video is updated only once every few seconds, and a hybrid scheme where full-motion video is transmitted when the system detects that a user is making a gesture and low-frame-rate video is transmitted at all other times. We studied people using our system for small-group discussions and found that the gesture-sensitive scheme was as effective for floor control as using full-motion video while requiring only a fraction of the bandwidth.

We begin by describing approaches to low-bandwidth videoconferencing and studies on the minimum frame rate necessary for effective communication. Next, section 3 describes the implementation of our gesture-sensitive conferencing system. Then, section 4 describes our user study and the findings. We conclude with a discussion of our results.

## 2. RELATED WORK

The required network bandwidth for videoconferencing can be lowered by using more efficient compression algorithms or by reducing the frame rate.

### 2.1. Low-bandwidth Video Compression

Discrete cosine transform (DCT) is used in most videoconferencing systems. A modern DCT compressor requires roughly 100 Kbps for a 320x240x15 fps video of a person's upper body [5]. If a DCT compressor is used below its target data rate, the video image may contain blocking artifacts and motions may appear jerky.

Two alternative approaches to DCT have been developed for extremely low bandwidth videoconferencing. The first approach encodes only the outlines of an image, the second approach encodes parameters to animate a 3D model of a person's head. Studies have shown that people can recognize the identity and facial expression of a person by the outlines of facial features [3][23]; thus, a colored image can be quantized into a binary image and only the edges in the binary image need to be encoded. A modern implementation of this idea delivers usable video at less than 10 Kbps [16].

The second approach analyzes a person's facial movements, transmits a description of the movements, and animates a 3D graphics model of the person's head at the remote end. The MPEG committee is standardizing this approach [27] and a modern implementation delivers usable video at less than 1 Kbps [8]. A drawback of this approach is that the animated person may not look natural since it is difficult to capture every nuance of the person's facial expression.

The DCT, the feature-outline, and the model-animation approach to video encoding do not use gesture information; thus, these approaches may be combined with our gesture-sensitive algorithm to achieve even lower data rates.

### 2.2. Minimum Required Frame Rate

The required network bandwidth can be lowered also by lowering the frame rate. The Portholes project has demonstrated that a frame rate as low as one update every five minutes can provide awareness in a work environment [7]; however, a direct application of this idea to remote classrooms may not be sufficient. Students often signal the desire to speak by raising their hands; this signal would be excessively delayed if transmitted through a Porthole-like system and the delayed delivery of floor control signals may disturb the instructional dialogue [14]. We augment a Porthole-like system to transmit floor control signals without delay.

Results of user ratings suggest that 5 fps is a lower bound on the acceptable frame rate. Tang and Isaacs reported that people rated 5 fps as tolerable [24]. Watson and Sasse found that audio and video is not perceived as synchronized at less than 5 fps [25].

Studies of user behavior found little difference in task outcome or communication behavior when the frame rate is lowered from 25 fps to 5 fps. Masoodian et al. studied pairs of people solving a jigsaw puzzle via a 5 and a 25 fps videoconferencing system and found that the frame rate had no effect on task completion time, number of utterances, amount of overlapping speech, number of speaker changes, or number of floor change attempts [18]. Jackson et al. studied pairs and groups of four people creating a tourist poster via a 5 and 25 fps videoconferencing system [13]. They found that the frame rate had no effect on the quality of the poster or the number of words spoken; however, they did find a small increase in the number of speaker changes when two people conferenced at 25 fps.

Experiments have also shown that lowering the frame rate from 25 to 15 and 5 fps does not decrease a person's understanding of the content of the video [10]. In fact, comprehension sometimes increased at 5 fps.

Studies reviewed so far suggest that 5 fps may be the minimum required frame rate; however, experiments have also shown that video can be useful at 1 fps. For example, novices were able to learn American Sign Language (ASL) at 1 fps and once they had learned it, they were able to recognize ASL gestures as effectively at 1 fps as at 30 fps [15].

All studies reviewed so far examined the effect of constant-frame-rate conditions, while our study examined the effect of non-uniform-frame-rate conditions.

## 3. DESIGN OF A GESTURE-SENSITIVE VIDEOCONFERENCING SYSTEM

We implemented a multiparty videoconferencing system that allows dozens of students to take a class from different locations. Each student, as well as the instructor, attends the class via a personal computer. Figure 1 shows the user interface. Note that all participants in the class are shown in a video grid. The usage model is that all participants can be seen and heard at all times.

Section 3.1 explains why we use reliable transmission for low-frame-rate video, section 3.2 describes a robust gesture-detection algorithm, and section 3.3 compares the required network bandwidth at different frame rates. The implementation framework is described in [5].

### 3.1. Reliable Streaming of Low-update Video

Most videoconferencing systems use user datagram protocol (UDP) for streaming full-motion video. UDP does not guarantee successful data delivery; however, since full-motion video has strong frame-to-frame coherence, an occasional frame dropping due to unsuccessful data delivery may not be noticeable [6].

Our system uses reliable UDP for streaming low-update video. We implemented a user level transmission control protocol (TCP) on top of UDP. This implementation was necessary since the native TCP's congestion control may unnecessarily limit transmission bandwidth.

We use reliable UDP for three reasons. First, users may not perceive the additional latency incurred by data retransmission since the network latency is a small fraction of the total latency if the video is updated once every few seconds.

Second, the failed transmission of a single frame may cause confusion in a classroom. For example, from our user study, we found that a raised hand does not stay up for more than a few seconds. If the video update time is comparable to the average time a hand remains raised, then a raised hand may be contained only in a single video frame. Suppose a student raises his hand and this frame is delivered to all the students but not to the instructor. The instructor, not seeing the raised hand, continues to lecture; the students not knowing of the failed transmission may think that the instructor is unresponsive, and the student with the raised hand may feel ignored.

Lastly, reliable UDP saves bandwidth in low-update video. Note in Figure 3a that I-Frames (compressed frames that reference only the current frame) are roughly ten times larger than P-Frames (compressed frames that reference previous frames). If an I-Frame is lost or corrupted, the following P-Frames will be incorrectly decoded, typically resulting in a "ghostly" image; thus, I-Frames are sent every few seconds to minimize error propagation. Low-update UDP streaming may send a frame once every few seconds, and a significant number of these frames should be I-Frames to limit the time an incorrect frame is displayed. If the streaming framework guarantees delivery, only the first frame needs to be an I-Frame.


Figure 1. Screen shot of our multiparty videoconferencing user interface.

### 3.2. Gesture Detection Algorithm

Within the computer vision community, the goal of using computers to detect, identify, and interpret human behavior has become a central research topic [20]. A review of the state-of-the-art in gesture tracking and recognition algorithms can be found in [9][19]. These algorithms are often designed with a limited assumption about the scene so that they can be useful for a wide range of applications; furthermore, the algorithms must minimize both false positive and false negative identifications. We use two assumptions to make our algorithm robust while the required computation is minimized. First, we assume that each camera will see a head and shoulder view of a single person. Further, we assume that the background behind any one person will not undergo rapid changes most of the time since the person is attending a class. This assumption allows us to detect hand motion using only motion cues instead of using both motion and color cues. Algorithms that use both motion and color cues often do not run in realtime [19]. Second, our algorithm only needs to minimize false negative identifications since the penalty for a false positive identification is a modest increase in bandwidth. This assumption simplifies the selection of threshold values in our gesture-detection algorithm and, consequently, allows us to use relatively coarse-grained computer vision processing to minimize computational load.

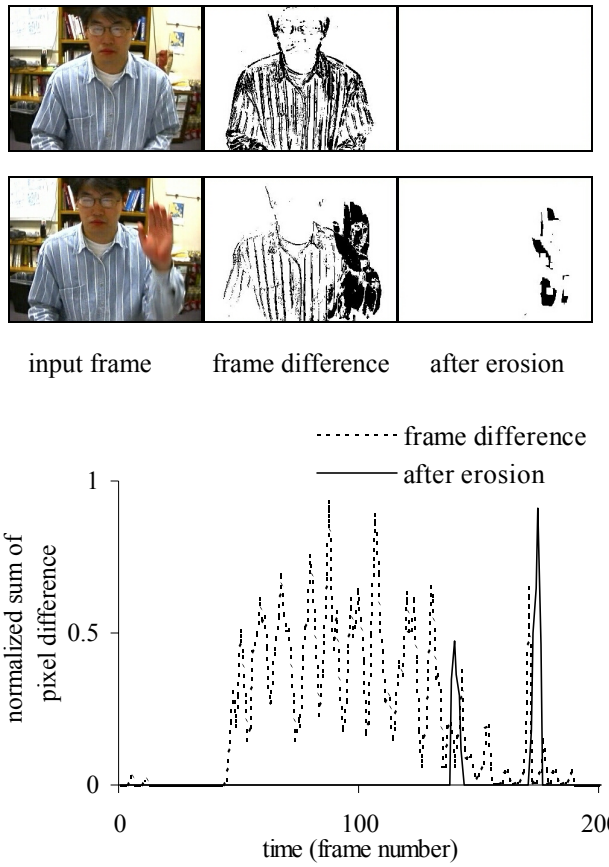
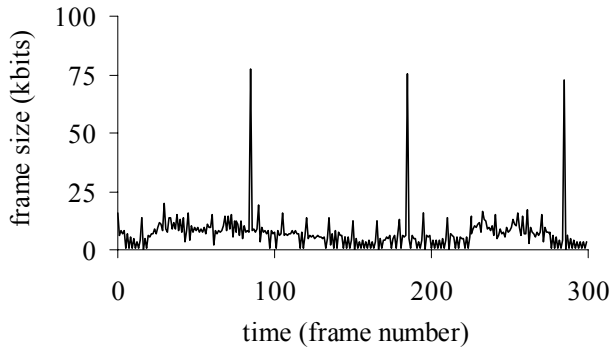
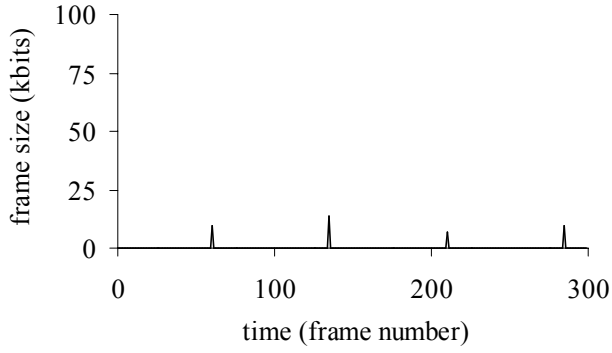



Figure 2. Gesture-detection algorithm. The top row of images shows an input frame, the pixel-by-pixel difference of this frame with respect to the previous frame, and the pixel difference after the erosion filter is applied. The second row of images shows the same processing pipeline when a hand is raised. The graph shows the frame-by-frame value of the sum of the frame difference and the sum of the eroded frame difference for a representative video. The person was initially sitting very quietly, next he moved back and forth in his chair, and finally he raised and then dropped his hand.

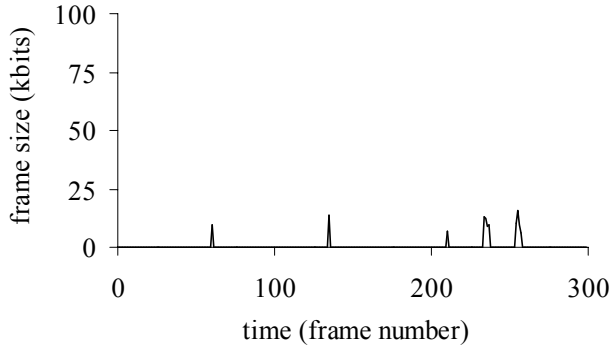
Figure 2 illustrates our gesture-detection algorithm. For each video frame, a video analysis module computes the pixel-by-pixel difference between the input frame and the previous frame. Next, an erosion filter is applied to the pixel difference. The erosion filter sets each pixel to the minimum value of that pixel and its eight neighbors. The effect of the erosion filter is to remove spurious pixels such as those from noise and to thin out the difference between the two frames. The erosion filter is applied four times, a number that we empirically determined to give good results. Note in Figure 2 that when a person slightly changes body position, any frame difference is essentially gone after erosion, whereas the erosion filter does not erase




a) full-motion videoconferencing




b) low-update videoconferencing




c) gesture-sensitive videoconferencing

Figure 3. Size of compressed frames for full-motion, low-update, and gesture-sensitive videoconferencing. All frames are 320 by 240 pixels and compressed using Microsoft's MPEG4 codec.

the motion of a hand being raised. Finally, the module sums the pixel values of the eroded frame and if this value exceeds a threshold, this frame is compressed and transmitted.

The algorithm just outlined cannot distinguish large body movements from hand motion since the erosion filter

may not filter out all body movements. Our usage model assumes that each camera will capture a head and shoulder shot of a single person; we use this a priori information to distinguish types of motion in the eroded frame. Observation of the eroded frame has shown us that hand motion typically causes a concentrated pixel difference in a single region while large body motion causes the eroded frame to show a pixel difference in many regions scattered over a larger area. For the eroded frame difference, the area of the bounding box containing non-zero pixel difference is computed, and only when this area is less than a threshold will it be considered as a possible hand motion. We have tested this algorithm for a variety of different users and under a variety of lighting conditions and have found it to be robust.

Figure 2 shows that a hand raise or a hand drop causes a spike in the graph of the eroded frame difference. We have used this characteristic to implement an ultra-low-bandwidth conferencing system that conveys only hand raises and hand drops. In this mode, instead of transmitting at full-motion whenever hand motion is detected, a frame is transmitted only at the end of each spike in the eroded frame difference. We did not investigate this ultra-low-bandwidth mode in our user study.

We implemented our gesture-detection algorithm using Intel's Image Processing Library [26]. The algorithm uses 15% of the processor cycles of a Pentium III 500MHz to process a 320 by 240 pixel video stream at 15 frames per second.

### 3.3. Effect of Frame Rate on Bandwidth

Figure 3 shows the network bandwidth required for full-motion video at 15 frames per second, low-update video at 1 frame every 5 seconds, and gesture-sensitive video. The graph plots the network packet size of each frame for a 20-second video sequence compressed using Microsoft's MPEG4 codec. The video sequence was representative of the videos recorded in our user study.

The three large spikes in the full-motion condition correspond to I-Frames. Low-update and gesture-sensitive conditions use reliable transmission; thus, the compression module does not generate any I-Frames after the initial I-Frame. In the gesture-sensitive condition, the first spike around frame 250 corresponds to a hand being raised and the following spike corresponds to the dropping of the hand. The largest compressed image for the full-motion, the low-update, and the gesture-sensitive condition are 77 Kbits, 14 Kbits, and 17 Kbits, respectively. The average bandwidths for the full-motion, the low-update, and the gesture-sensitive condition are 108 Kbps, 2 Kbps, and 11 Kbps, respectively.

The actual bandwidth requirement of gesture-sensitive conferencing will depend on the frequency of hand raising and other gestures. From our user study, we found that one hand being raised every 20 seconds per person would result in a very lively discussion environment; thus, the expected bandwidth in practice should still be significantly less than that required for full-motion video.

## 4. USER STUDY ON THE IMPACT OF FRAME RATE ON BEHAVIOR

The goal of this user study is to evaluate the impact of frame rate on conversational behavior, specifically, people's ability to request to speak and to judge when to stop speaking in a remote classroom environment.

### 4.1. Methodology

We used the task of group discussion. To suppress the effect of subjects' background knowledge, we chose a simple topic to stimulate lively discussions. The discussion scenario was that a successful software engineer in her late twenties had recently been laid off. Having worked hard since graduating, she wants to take a year off to travel. She would prefer not to spend more than twenty-five thousand dollars. The discussion topic was where she should go, what she should do, and how she should do it frugally.

Eight groups of four people per group participated in the discussion. The participants were current and recent graduates of Stanford University. The people in each group first met face to face in our lab and then each sat in front of a computer and continued the conversation using our videoconferencing software. The participants were told that they should raise their hand to indicate a desire to speak and that they should be called on before speaking. The last person who spoke chose the next speaker. The hand raising protocol was designed to create a polite but lively discussion environment.

The three experimental conditions were full-motion at 15 frames per second, low-update at 1 frame every 5 seconds, and gesture-sensitive, where automatically detected gestures were transmitted at full-motion and at all other times frames were transmitted as in the low-update condition. In a pilot user study, we also tested low-update conditions at 1 frame every 5 minutes, as in the Portholes system [7], and at 1 frame every 10 seconds; however, users considered these frame updates as too infrequent to be worth paying attention to. We did not try updates at a rate higher than 1 frame every 5 seconds so the low-update condition would have difficulties conveying gestures and facial expressions. In summary, the full-motion condition conveys facial expressions, gestures, and posture positions, the gesture-sensitive condition conveys gestures and posture positions, and the low-update condition conveys posture positions.

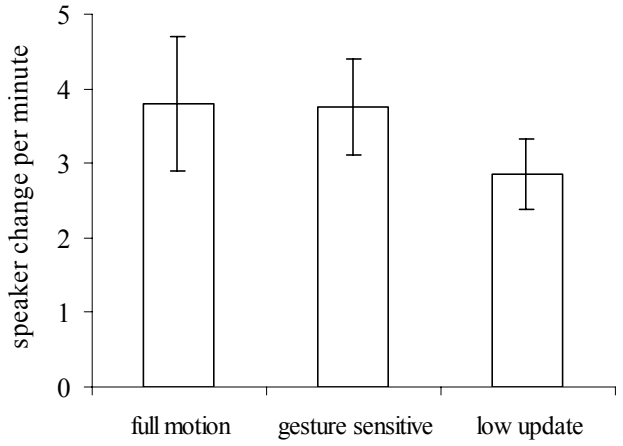



Figure 4. Average number of speaker change per minute during the discussion.

Each video frame had a resolution of 320 by 240 pixels and was captured using a LogiTech QuickCam Pro 3000 USB camera. We used 20-inch monitors and set the display resolution at 640 by 480 pixels, so the videos of the four participants covered the entire screen. For each of the three conditions, a three-minute warm up preceded five minutes of discussion. Each group held discussion using all three conditions, and the order of the conditions was counterbalanced.

The audio and video of each participant were recorded using our software. After the discussion, participants filled out a questionnaire and were interviewed to collect open-ended feedback.

## 4.2. Results

A measure of the liveliness of a discussion is the number of speaker changes. Figure 4 shows the average number of speaker changes per minute during the discussion for the three frame-rate conditions. The low-update condition resulted in fewer speaker changes than the full-motion condition, while the gesture-sensitive condition achieved a similar number of speaker changes.

Figure 5 shows the results of the questionnaire. Table 1 lists the survey questions. Note that the gesture-sensitive condition was more effective in supporting floor control than the low-update condition. Questions on engagement and enjoyment showed less difference between the three conditions, indicating perhaps that these awareness metrics are less sensitive to frame rate. Overall, the gesture-sensitive and the full-motion condition were judged to be useful to the discussion while the low-update condition was not.

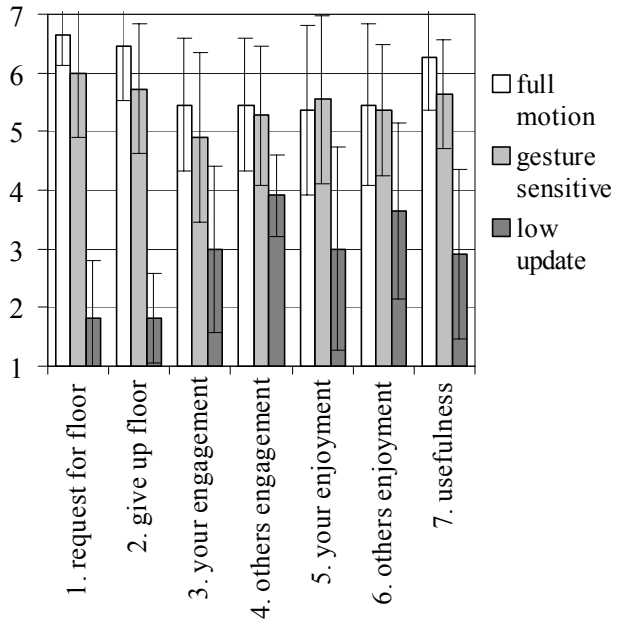



Figure 5. Survey results. Graph shows the average of users' responses to the statements in Table 1. A response of 1 corresponds to strongly disagree, 4 corresponds to neutral, and 7 corresponds to strongly agree.

| | |
|---|---|
| Floor control | 1. This condition did not limit your ability to request for floor (signal your desire to speak) |
| | 2. This condition did not limit your ability to judge when others want to speak |
| Engagement and enjoyment | 3. You were engaged (absorbed) in the discussion under this condition |
| | 4. Other people were engaged (absorbed) in the discussion under this condition |
| | 5. You enjoyed the discussion under this condition |
| | 6. Other people enjoyed the discussion under this condition |
| Utility | 7. Overall this condition was useful for the discussion |

Table 1. Survey questions posed to the users.

## 5. DISCUSSION

Figure 5 shows that participants viewed the low-update condition as ineffective for floor control, and yet Figure 4 shows that the three frame rate conditions did not result in as large a difference in the speaker changes as Figure 5 might suggest. To explain this finding, we define two terms, floor holding time and floor change latency. The floor holding time is the time between speaker changes, which is about 20 seconds on average in our user study. The floor change latency is the time between when a person requests to speak and when that person begins to speak. The floor change latency introduced by the video medium is on average 2.5 seconds for the low-update condition and 33 milliseconds for the full-motion and the gesture-sensitive condition. Since the additional latency introduced by the low-update condition is a small percentage of the floor holding time, we should not see a large decrease in the number of speaker changes even though participants felt that the low-update condition was ineffective for floor control.

If the frame rate in the low-update condition were decreased to the order of the floor holding time, then we would expect a large decrease in the number of speaker changes. On the other hand, if the floor holding time were significantly longer than 20 seconds, as is the case in a more formal discussion or lecture, then we may not be able to measure any difference between the three conditions in terms of speaker change.

A common complaint about the low-update and the gesture-sensitive condition is that people can be caught at a moment that makes them look silly, typically in the middle of a movement, with the consequence of all the participants laughing. This effect may be minimized if the time when the camera will take the next shot can be indicated to the user, perhaps by a graphical count-down indicator.

Hand raising is the predominant social protocol for requesting to speak in a classroom, but it is not always required for effective floor control. When the participants know each other well, they learn to thread their comments or questions between the natural breaks in the current speaker's utterance; thus audio conferencing alone may suffice under these conditions. In a pilot study, we asked groups of four people who knew each other well to participate in our study. Unlike our main study, these participants were not told of any hand-raising protocol. We found that the participants often did not look at the videos. They would start to speak as soon as the current speaker pauses.

Even when participants always raise their hands to request to speak, floor control also depends on other signals. An experienced speaker, for example, monitors the listeners' gaze, facial expressions, and body positions. If listeners appear to be confused, the speaker may pause and invite a question or comment from the audience. Unlike full-motion videoconferencing, gesture-sensitive conferencing is ineffective at transmitting gaze, facial expressions, and high-frequency body movements. One way to minimize this shortcoming is to detect the natural pauses in a speaker's delivery and stream at full-motion during these pauses. Speakers tend to not look at the listeners during an utterance but to look at them at the end of the utterance [2], presumably to check for feedback from the audience. Streaming during the speech pauses may offer enough visual feedback to allow a speaker to adapt to the audience. We plan to conduct user studies to verify or refute this speculation.

Instead of gesture-sensitive conferencing, an alternative improvement to low-frame-rate conferencing is to allow students to use a keyboard to signal the desire to speak, for instance by overlaying the student's image with an iconic representation of a raised hand. However, instructors in a face-to-face classroom expect a spectrum of different gestures, from the hesitantly raised hand to the must-speak-immediately thrust. It is unclear how to effectively map different gestures to graphical icons and whether such a system will be easy to learn and use. Given that congenitally blind children make gestures similar to those of sighted children, even when they know the listener is blind [12], the ability to make and interpret gestures may be inborn; thus, a system that conveys gestures in their natural form may have a biological performance advantage.

## 6. CONCLUSION

Multiparty videoconferencing with even a small number of people is often infeasible due to the high network bandwidth required. Commodity videoconferencing products often compete in the maximum visual fidelity of a single video stream that they can deliver. Our research explores the minimum visual fidelity necessary for videoconferencing to be effective. Our contributions are (1) the design and implementation of a gesture-sensitive videoconferencing system and (2) a user study on the effect of frame rate on small-group discussions in a remote classroom environment.

The three frame-rate are (i) full-motion, which conveys facial expressions, gestures, and postures, (ii) gesture-sensitive, which conveys gestures and postures, and (iii) low-update, which conveys postures. Our data suggests that conveying postures alone is insufficient for small group discussions due to difficulties with floor control. Our data also suggests that conveying gestures in addition to postures is a viable option if limited bandwidth would otherwise prevent using videoconferencing at all.

Our gesture-sensitive videoconferencing software is available for download at [29]. For future work, we plan to incorporate more sophisticated computer vision modules to detect head and eye movements so that these signals can also be selectively transmitted.

## 7. ACKNOWLEDGMENTS

This project is sponsored by the Stanford Interactive Workspace Project, the Stanford Immersive Television Project, and the Stanford Learning Lab. We wish to thank Jingli Wang for participating and improving the pilot user study, and Evelin Sullivan, Matthew Eldridge, Albert Huntington, Linda Wang, and the reviewers for critiquing and improving the draft.

## 8. REFERENCES

[1] A. Anderson, A. Newlands, J. Mulin, A. Fleming, G. Doherty-Sneddon, and J. Velden. Impact of Video-Mediated Communication on Simulated Service Encounters. *Interacting with Computers*, pages 193-206, 1996.

[2] M. Argyle and M. Cook. *Gaze and Mutual Gaze.* Cambridge University Press, 1976.

[3] V. Bruce. The Role of the Face in Communication: Implications for Videophone Design. *Interacting with Computers*, pages 166-176, 1996.

[4] C. Chen, M. Rouan, J. Edwards, and C. Moore. An Exploratory Study on the Impact of Broadcast Courses at Stanford University. Prepared for the Stanford Dean of Engineering, 2000.

[5] M. Chen. Design of a Virtual Auditorium. *Proceedings of ACM Multimedia*, pages 19-28, 2001.

[6] M. Claypool and J. Tanner. The Effects of Jitter on the Perceptual Quality of Video. *Proceedings of ACM Multimedia*, pages 115-118, 1999.

[7] P. Dourish and S. Bly. Portholes: Supporting Awareness in a Distributed Work Group. *Proceedings of CHI*, pages 541-547, 1992.

[8] P. Eisert and B. Girod. Analyzing Facial Expressions for Virtual Conferencing. *IEEE Computer Graphics & Applications:* Special Issue on Computer Animation for Virtual Humans, pages 70-78, 1998.

[9] D. Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, pages 82-98, January 1999.

[10] G. Ghinea and J. Thomas. QoS Impact of User Perception and Understanding of Multimedia Video Clips. *Proceedings of ACM Multimedia*, pages 49-54, 1998.

[11] E. Isaacs, T. Morris, T. Rodriguez, and J. Tang. A Comparison of Face-to-face and Distributed Presentations. *Proceedings of CHI*, pages 354-361, 1995.

[12] J. Iverson and S. Goldin-Meadow. Why people gesture when they speak. *Nature*, volume 396, pages 228, November 1998.

[13] M. Jackson, A. Anderson, R. McEwan, and J. Mullin. Impact of Video Frame Rate on Communicative Behavior in Two and Four Party Groups. *Proceedings of CSCW*, pages 11-20, 2000.

[14] G. Jancke, J. Grudin, and A. Gupta. Presenting to Local and Remote Audiences: Design and Use of the TELEP System. *Proceedings of CHI*, pages 384-391, 2000.

[15] B. Johnson and J. Caird. The Effect of Frame Rate and Video Information Redundancy on the Perceptual Learning of American Sign Language Gestures. *Proceedings of CHI*, pages 121-122, 1996.

[16] J. Li, G. Chen, J. Xu, Y. Wang, H. Zhou, K. Yu, K. Ng, and H. Shum. Bi-level Video: Video Communication at Very Low Bit Rates. *Proceedings of ACM Multimedia*, pages 392-400, 2001.

[17] R. Malpani and L.Rowe Floor Control for Large-Scale Mbone Seminars. *Proceedings of ACM Multimedia*, pages 155-163, 1997.

[18] M. Masoodian, M. Apperley, and L. Frederickson. Video Support for Shared Work-space Interaction: an empirical study. *Interacting with computers*, pages 237-273, 1995.

[19] V. Pavlovic, R. Sharma, and T. Huang. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pages 677-695, July 1997.

[20] A. Pentland. Looking at People: Sensing for Ubiquitous and Wearable Computing. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pages 107-118, January 2000.

[21] L. Rowe. ACM Multimedia Tutorial on Distance Learning, 2000.

[22] J. Short, E. Williams, and B. Christie. *The Social Psychology of Telecommunications*. London, U.K. Wiley, 1976.

[23] B. Stapley. Visual Enhancement of Telephone Conversations. Ph.D. Thesis, University of London, 1972.

[24] J. Tang and E. Isaacs. Why Do Users Like Video? Studies of Multimedia-Supported Collaboration. *Computer-Supported Cooperative Work: An International Journal*, pages 163-196, 1993.

[25] A. Watson and A. Sasse. Evaluating Audio and Video Quality in Low-cost Multimedia Conferencing Systems. *Interacting with Computers*, pages 255-275, 1996.

[26] Intel Image Processing Library http://developer.intel.com/software/products/perflib/ipl

[27] MPEG-4 Overview. ISO/IEC JTC1/SC29/WG11 N4030, March 2001.

[28] Stanford Center for Professional Development. http://scpd.stanford.edu/scpd/about/history.htm

[29] Stanford Video Auditorium http://graphics.stanford.edu/~miltchen/VideoAuditorium