A Robust Abstraction for First-Person Video Streaming: Techniques, Applications, and Experiments

Neil J. McCurdy[1], William G. Griswold[1], and Leslie A. Lenert[2]

[1]Department of Computer Science and Engineering
University of California, San Diego, La Jolla, CA

[2]Veterans Affairs San Diego Healthcare Systems, San Diego, CA

{nemccurd,wgg,llenert}@ucsd.edu

Abstract

The emergence of personal mobile computing and ubiquitous wireless networks enables powerful field applications of video streaming, such as vision-enabled command centers for hazardous materials response. However, experience has repeatedly demonstrated both the fragility of the wireless networks and the insatiable demand for higher resolution and more video streams. In the wild, even the best streaming video mechanisms result in low-resolution, low-frame-rate video, in part because the motion of first-person mobile video (e.g., via a head-mounted camera) decimates temporal (inter-frame) compression. We introduce a visualization technique for displaying low-bit-rate first-person video that maintains the benefits of high resolution, while minimizing the problems typically associated with low frame rates. This technique has the unexpected benefit of eliminating the "Blaire Witch Project" effect – the nausea-inducing jumpiness typical of first-person video. We explore the features and benefits of the technique through both a field study involving hazardous waste disposal and a lab study of side-by-side comparisons with alternate methods. The technique was praised as a possible command center tool, and some of the participants in the lab study preferred our low-bitrate encoding technique to the full-frame, high resolution video that was used as a control.

1. Introduction

The emergence of personal mobile computing and ubiquitous wireless networks allows for remote observation in uncontrolled settings. Remote observation is powerful in situations in which it is not possible or too dangerous for an observer to be present at the activity of interest. These include coverage of breaking news, emergency response, or grand parents joining the grandchildren on a trip to the zoo. The application investigated in this paper is video-support for a supervisor overseeing hazardous materials disposal.

Despite incredible advances in wireless networking and the mobile devices connected by it, our repeated experience is that wireless networks in uncontrolled settings are fragile, and there is seemingly unlimited demand for more video streams at higher resolution. Modern video streaming techniques heavily depend on temporal (inter-frame) compression to achieve higher frame rates, while minimizing the impact on resolution when operating at the network's capacity. Unfortunately, the panning motions common to first-person mobile video (captured from a headcam, say) virtually eliminates inter-frame compression. To stay within the available bandwidth, either the frame rate or the resolution must be reduced. In applications like hazardous materials disposal, image resolution cannot be sacrificed, making a low-frame-rate encoding the only viable option. Ironically, lower frame rates further reduce the likely overlap between frames, further reducing inter-frame compression.

The problem with low-frame-rate video is that a one-second interval between frames is long enough to disorient the viewer. This is especially true with head-mounted cameras because it may only take a fraction of a second for the view to rotate 180 degrees. With little or no overlap between successive frames, the viewer lacks the information required to understand how the frames relate to one another.

In this paper, we present a visualization technique that minimizes the confusion caused by low-frame-rate video, using modest hardware and processing. If the orientation of the camera is known – either by attaching tilt sensors and an electronic compass to the cameras, or by using an online vision processing algorithm on the cameras – we can generate a visualization that shows the viewer how con-

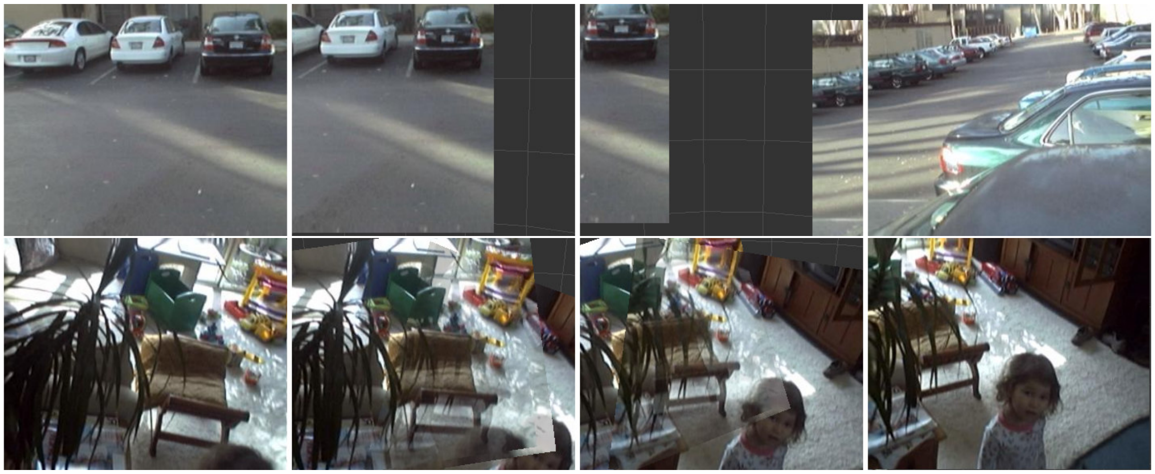

Figure 1. *Snapshots of two transitions in progress. The top row depicts a camera pan from left to right where the frames do not overlap. The bottom is a morph from the frame on the left to the frame on the right as the camera pans down and to the right to look at the child. The live experience is one of smooth camera movement.*

secutive frames relate to one another. The visualization takes the form of a dynamic *transition* similar to those described for switching between two streaming cameras located in the same environment [9]. A transition (Fig. 1) has two components: movement (rotation) of the viewing portal from one frame to the next, and a gradual alpha-blend between the overlapping portions of the frames. If the frames do not overlap at all, the current frame rotates off of the screen, a spherical grid (as viewed from the center of the sphere) continues to show the degree and direction of orientation, and finally the next frame rotates onto the screen. The net effect is a high-frame-rate interpolation of the camera's motion between the frames. These transitions intuitively convey the relative positions of the frames, and no users in our user study reported anything more than occasional temporary confusion when watching long sequences of these transitions. Due to the visual nature of this work, we encourage the reader to view short video clips of transitions downloadable from the web (`http://ubivideos.homeip.net`).

We explore the features of this approach in part with a field study of a hazardous materials (hazmat) supervisor remotely monitoring a live video feed – transmitted over a "broadband" cellular network – of two hazmat workers disposing of hazardous chemicals. The camera was mounted on the mask of one of his team members. Such a system configuration is motivated by a response in a damaged and chaotic environment. The supervisor's impressions of our visualization technique were surprisingly favorable, and he dismissed the alternative encodings that were available. The unmodified low-frame rate video left him feeling disoriented, and the low-quality 5fps (frames-per-second) video was so choppy and disorienting that it interfered with his thinking and made him nauseated.

We explore the finer distinctions among the various approaches to low-frame-rate video with a laboratory study in which 14 subjects were asked to view video clips of three different scenes that were encoded in four different ways. A surprising result of this study is that four of the subjects actually preferred watching our 1fps transition-enhanced video over full-frame (12fps), high quality video. Nearly all of the participants preferred our visualization to the 5fps video clip that was encoded at a comparable bitrate. One further interesting result is that nearly all of the participants were unable to discern the difference between a clip that performed a simple alignment and blending between frames, and one that also performed a morph between the frames to produce more seamless transitions. This result can be explained by the brain's ability to commit *closure* with minimal cognitive load when modest amounts of visual information are missing [8].

The remainder of this paper is organized as follows: In section 2 we motivate the use of video in a disaster response setting, and describe the constraints that such an environment places on technical solutions. In section 3 we describe our solution, and in section 4 we discuss related work. Sections 5 and 6 present our field and lab studies.

2. Motivation

There are many situations in which high-panning low-bit-rate video can have value. Consider, for example, CNN coverage of hurricanes or remote war-torn areas where CNN resorts to satellite-phone video segments. These feeds are tolerable for the talking-head shots, but panning of the surrounding environment to show viewers what is happening results in a dissatisfying choppy, grainy image. There are also man-on-the-street news reporting scenarios where it might be desirable to look at low-bitrate video. Breaking news such as an accident, prior to the arrival of traditional television cameras, could be viewed through citizen cameras with feeds transmitted over cellular networks, or overlooked news could be streamed direct to the internet by citizen mobile phones.

Hazardous Materials Cleanup. The use of video during the early stages of a disaster response, or even during the late stages of a chemical clean-up are scenarios that can be impacted today. This paper focuses on this latter scenario, and we have used the requirements of a hazardous materials (hazmat) supervisor as the requirements for our video streaming solution. We consulted with the Hazardous Materials Business Plan Manager (hereafter referred to as Tod) at the University of California, San Diego (UCSD) to determine how live video might be used at a hazmat scene.

As a supervisor, it is Tod's job to know what is going on, to interface with the various entities on scene (such as fire fighters, witnesses, and lab managers), and to supervise the stabilization and cleanup of the environment. Live video feeds from the scene would help Tod assess the health and safety of his team, aid in identifying hazards, and allow experts outside of the hotzone to assist with operations.

Networking Challenges. The significant radio interference at a disaster scene (both natural and man-made) wreak havoc with communication. In contrast to existing hazmat video transmission systems which typically use analog signals, we have decided to use a digital signal for a number of reasons. First, in the larger-scaled deployment of our parent system, RealityFlythrough, we are piggy-backing on a state-of-the-art wireless mesh network [2] that is deployed by first-responders to support the coordination of medical treatment for victims. Second, the varying conditions of the network caused by radio noise can be better managed in the digital domain. Frame rates can be throttled and image quality can be degraded in a controlled manner. Most importantly, we can guarantee eventual delivery of error-free images (with a very high latency) when conditions are so bad that only a small amount of data can trickle through the network. And third, we can use the same bandwidth managing techniques to support multiple cameras.

Radio interference in a digital mesh network results in frequent disconnects and low throughput. 802.11b has an expected bitrate of 6.4Kbps, but noise, overhead introduced by the mesh network, and the many other clients competing for bandwidth have reduced the effective bandwidth to 100Kbps for each camera in a typical 3-camera deployment. Similar conditions are found in an alternate deployment scenario which uses a cellular network instead. In this case, immature technology is the main source of fragility.

Video Compression Challenges. The conditions that have been outlined so far present a significant challenge for video compression. The video stream produced by a head-mounted camera is typically high-panning due to the natural head movements of the wearer. High-panning video usually has very little redundancy between frames, rendering traditional codecs that rely on temporal redundancy ineffective. With low temporal redundancy in the video input, most codecs do little better than motion JPEG (MJPEG) which simply performs spatial compression on each frame in the video sequence.

In the heavily constrained networks described above, where the frame rate must be reduced to maintain image quality, the increased interval between frames further reduces temporal redundancy, minimizing the bitrate savings of the decreased frame rate. The result is a heavily decimated frame rate. As the frame rate drops, it becomes difficult to track objects, and eventually it is even difficult to orient yourself in the scene.

Traditionally, the only option at this point has been to reduce the image quality to increase the frame rate to non-disorienting levels. Our approach preserves image quality while mitigating the negative effects of low frame rates.

3. Our Approach

To reduce the disorienting effects of low-frame-rate video, our concept is to perform a dynamic visual interpolation between frames using meta data captured from a digital pan/tilt compass or inferred using vision techniques. In particular, we align the frames in a spatially consistent way in a 3d graphics environment, and then use rotational and translational motion to segue between the frames, producing a high-frame-rate experience that captures the effects of camera motion. Because precise frame stitching is impossible in real-time using 2D data, we use a dynamic crossover alpha-blend to help the viewer correlate the information in the overlapping parts of the frames.

An imperfect alignment between two frames, due to, say, inaccurate sensor readings is less of an issue than might be expected. *Closure* is a property of the human visual system that describes the brain's ability to fill in gaps when given incomplete information [8]. It is a constant in our lives; closure, for example, conceals from us the blind spots that are present in all of our eyes. So while there is ghosting, and maybe even significant misregistration between frames, the human brain easily resolves these ambiguities.

The rest of this sections describes the details of our approach.

Creating a Panoramic Effect. Our approach can be described as the creation of a dynamically changing and continually resetting spherical panorama. Each incoming frame is positioned on the panorama, and projected onto a plane that is tangential to the sphere to avoid distortion. A dynamic *transition* then moves the user's viewpoint from the current position within the panorama to the incoming frame's position (Fig. 1, bottom). The user's viewport has the same field of view as the source camera, so the frame fills the entire window once the transition is complete. Movement between frames looks like smooth camera panning. There may also be a translational (shifting) motion effect if the camera moves forward or backward through the scene.

A new panorama is started when consecutive frames do not overlap (Fig. 1, top). The frames are positioned at their relative locations on the sphere, with an appropriate gap between them. To help the user stay oriented, a wireframe of the sphere that serves as the projective surface is displayed. Horizontal and vertical rotations are thus easily recognized. The grid wireframe could be further enhanced by including markers for the equator and the cardinal directions.

The planar simplification of 3d space only works for a short interval when cameras are mobile. For this reason, at most five frames are placed in a given panorama. The oldest frame is discarded when this limit is reached. This is not a significant compromise because the source and target frames of a transition mostly fill the viewport, and any other frames in the panorama are filling in around these two.

Frame placement in the panorama is managed through a robust two-level scheme, as described in the rest of this section.

Image-based Frame Placement. When inter-frame rotations are not too large, we use an implementation of Lowe's SIFT algorithm [7] to find matching points between a new frame and the previous frame, and then do a best-fit alignment of the frames to fit the new frame into the panorama. The point-matching is performed on the camera units in real-time, and the list of matched points between the current frame and the previous frame are transmitted with each frame. In order to perform the matches in real-time on our camera devices, the frame is downsampled to a quarter resolution (QCIF instead of CIF) prior to analysis by SIFT. The result is good even at this lower resolution.

Each new frame is aligned to the previous frame by determining an affine correspondence between the frames. We look at the relative position, orientation, and zooming based on the two matching points in each frame that are furthest apart. After aligning the new frame, the frame is warped so that the matching points are exactly aligned. Surrounding points are warped by an amount proportional to the inverse of the distance to the neighboring control points. A transition to this new frame thus involves a morph as well as the standard rotation and alpha-blend. At the end of the transition, the new frame will be unwarped, and all of the other frames will be rotated and warped to match the control points in the new frame.

Even with point matching, the alignment is not fully precise. Our planar simplification of 3d space makes objects in the scene that are at depths different to those of the points that have been matched be less accurately aligned. Even if the depths of the matching points were recovered, the number of matching points (10-20) is very small relative to the number of objects and object depths in the scene, so any recovered geometry would be coarse. Also, since we are operating in real environments, dynamic objects that move between frames will not have any point correspondences, and thus will not be accurately aligned. Nonetheless, closure helps this technique produce very pleasing results.

Sensor-based Frame Placement. When SIFT fails to produce matching points for a new frame, the frame's placement depends on sensor data gathered from the camera rig. The camera units we use are integrated with tilt sensors and electronic compasses that record the tilt, roll, and yaw of the cameras at 15hz. This information allows us to position the frames on the sphere. However, the sensor accuracy is not good enough for generating a multi-frame panorama. Thus, the placement of such a frame initiates a new panorama with the single frame. The rotational part of the transition is still performed with the dynamic alpha-blend, using the previous frame's and new frame's relative sensor data. However, since we do not have information about the relative or absolute locations of the frames, we are unable to determine the relative translational positioning between frames. The resulting experience mitigates the confusion caused by low frame-rate video, but often lacks the aesthetics of the panorama and higher precision placement.

4. Related Work

We are not aware of any related work that directly addresses the conditions we have set out to handle in this paper, but there is some work that handles subsets of these problems.

RealityFlythrough, which provides ubiquitous video support for multiple mobile cameras in an environment, uses visualization techniques similar to the one we propose in this paper, but for inter-camera transitions [9]. It requires knowledge of the positions of the cameras, as well as the orientations, limiting its use to environments where ubiquitous location sensors are available, such as outdoors.

Irani, et al. directly address the problem of encoding panning video [5]. They construct a photo mosaic of the scene, and are then able to efficiently encode new frames by using the difference between the frame and the mosaic. With this technique, it no longer matters if consecutive frames have much overlap because the assumption is that similar frames have overlapped enough in the past to construct the mosaic. Unfortunately, mosaic-based compression cannot be used in our scenario because our cameras are mobile. Mosaic-based compression works well as long as the camera remains relatively static and pans back and forth over the same scene, but if the camera moves through the scene, there will be little opportunity to find matches with previous images. Essentially mosaic-based compression extends the search window for similar frames. If there are only a few similar frames, it does not matter how big the search window is, as there will rarely be a match.

There are many examples of codecs that are designed to work in wireless, low-bit-rate environments, although these codecs generally rely on the significant temporal compression that is possible in "talking-head" video. H.264 [15] (also known has MPEG4-10) represents the current state-of-the-art. When compressing first-person-video at low bitrates, though, there is little perceptible difference between H.264 and the more common MPEG4-2 [1] (commonly referred to simply as MPEG4). This is not surprising considering the low temporal redundancy.

A non-traditional approach to video compression proposed by Komogortsev, varies the quality of the video based on where the viewer is looking [6]. By using eye-gaze-trackers on the viewer, and predicting where the viewer will look next, the overall image quality can be low, but the perceived quality would be high. This approach would be difficult to implement in our scenario because the network latency is so high (4-5 seconds) that the gaze direction would have to predicted far in advance.

There has been substantial work on generating panoramas from still photographs [13, 4]. Real-time dynamic creation of panaoramas on a handheld camera device has been used to help with the creation of a static panorama [3]. Panoramas can also be efficiently created from movie cameras assuming the camera's position is relatively static [12]. All of these techniques require some way to match points between images. We rely heavily on Lowe's SIFT algorithm [7], specifically the Autopano implementation of it (`http://http://autopano.kolor.com/`).

A technique to remove distortions during image morphs is described by Seitz and Dyer [11]. This technique produces natural morphs, but it requires manual user intervention with each morph and is thus not applicable in our scenario. In practice, our technique rarely produces morphs that might cause disorientation, so morphing improvements would only be an aesthetic luxury.

5. Hazmat Field Study

We had several goals for our field study. First, we wanted to know if our visualization technique was suitable for a hazmat command center. Second, we wanted to see if our system could work in a realistic environment for an extended period of time. And third, we wanted to discover the motion model of a head-mounted camera afixed to someone doing a real job, oblivious to the presence of the camera.

5.1. Experimental Setup

The Scene. Every week, two members of the the UCSD hazmat team perform a maintenance task that doubles as an training exercise for response to an accident. All of the hazardous waste that has been collected from labs around the university is sorted, and combined into large drums in a process that is called *bulking* of solvents. This task serves as an exercise, as well, because full hazmat gear must be worn during the procedure, giving the team members (we will call them *bulkers*) experience putting on, wearing, and performing labor-intensive tasks in gear that they will use at an incident site.

The Equipment. It was important to make the camera system as wearable and unobtrusive as possible, given our desire to discover the real motion models of the camera.

We attached a disassembled Logitech webcam (∼\$100) to the front of the mask, and sewed a tilt sensor manufactured by AOSI (∼\$600) into the netting of the mask that rested on the top of the head. These devices connected to a Sony Vaio U71P handtop computer (∼\$2000) which was placed in a small backpack. Tod, the team leader introduced in section 2, insists that the bulking experience be a replica of real-world hazmat scenarios, so we chose to transmit the video across the Verizon 1xEVDO network, which might be the only readily available network if, say, a burnt out lab were being cleaned up. The video feed was transmitted to our server, a standard VAIO laptop (FS-790P ∼\$1600) connected via 802.11 to the campus network.

The 1xEVDO upstream bitrate was measured at between 60 and 79Kbps, and the campus downstream bitrate at 3.71Mbps. We fixed the frame rate of the video feed to .5fps to ensure that we would stay within the range of the 1xEVDO upstream speed.

The Task. We had Tod use the video that was being transmitted by one of his bulkers to explain to us the bulking process. This think-aloud interaction is realistic in that Tod needs to train others in how to conduct his task for times when he is on vacation or out sick. For us, this interaction served several purposes: (1) It would give Tod a reason to be viewing the video, (2) it would encourage him to verbalize his impressions of the system, and, (3) it would allow us to observe the effectiveness of the video stream as a communicative device. Did the video provide enough detail to help illustrate what he was describing, and at a fundamental level, did he understand what was going on?

As an expert, Tod's subjective opinion of the system was important to us. The requirements are domain specific, and only someone who has experience operating in a command center can know if the quality of the video is appropriate for the task.

5.2. Results

Our camera system was worn by one bulker for the entire exercise which lasted for roughly 64 minutes. The bulking task was very demanding for our visualization system because the close quarters of the bulking environment limit the field of view, and the heavy physical activity creates drastic camera pans from ground to horizon.

Despite these challenges, Tod reacted favorably to the visualization. He was oriented immediately: "Ok, so this is following Sam as he's moving around the room. And as you can see they have a lot of work ahead of them." The image quality was good enough for him to identify the characteristics of chemicals: "This tells me a little bit about the viscosity. I can see the liquids, whether they're plugging up. Sometimes you get some chunks in there. And the thicker stuff – gels – looks like we had a little bit in there..."

Tod expressed an interest in flipping through the individual frames so that we could really study the pictures. We showed him how he could pause the feed and move back and forth through the images while still getting the benefit of the visualization. The visualizations helped us stay oriented as he was describing the process, and saved him from having to explain the relative positions of the images.

We then discussed how our visualization compared to the the normal view of low frame-rate data which at these speeds looked more like a sequence of still photos. "Literally for me, at the moment I would just go full screen on this particular moving one (our visualization)... I'm not really even paying attention to this one (the low frame-rate stream). The individual photos clicking through. I could be disoriented with that one... It would tell me that they're moving around, but after that it's not giving me anything that I really need for decisions."

Tod concluded with his assessment of the system: "Let me tell you what I like about it. It's not overwhelming. It's appropriate. It's not a huge distraction. That's one of the things you have to be concerned about – the level of distraction.... Yeah, I think you got it."

Tod also had recommendations for improvement: he would like to have multiple cameras so that he could see the scene from multiple angles, he requested wider-angle lenses, and he wondered if he could set up fixed cameras as well.

5.3. Followup

During the study we were of course unable to show Tod other possible encodings of the data. Thus, we returned a few days after the experiment and presented him with a re-creation of the experiment with 5fps video encoded at bitrates comparable to the original experiment, using FFMPEG's MPEG4 codec (`http://ffmpeg.sourceforge.net`). His reaction surprised us because we assumed that the necessary drop in video quality (resolution) would make the video unusable for hazmat: "I don't have a problem with the resolution on the right (the 5fps video), but it's almost flipping through so fast that you're not orienting yourself to what's going on... Yeah, I like the slower frame rate. It's not so much because of the resolution, it's the amount of time that it takes me to know what I'm looking at... [The 5fps video] is snapping too fast – it's too busy – it interferes with my thinking, literally, it's messing with my head."

Even after showing him the high quality 6.67fps feed that had been captured directly at the camera, Tod still thought our abstraction was more appropriate for a command center considering everything else that is going on. A command center needs to maintain a sense of calm [14]. "This is just one piece of information that you're going to be getting. The phone is going to be ringing, people are going to be giving you status reports. The [higher frame-rate video] is just too busy."

6. Lab Study

Intrigued by Tod's observations during the field study that our visualization method may actually be *more* pleasurable to watch than high fidelity first-person video, we increased the scope of our planned lab study to determine if our visualization method would have broader appeal. Might it actually be an alternative to the sometimes nauseating, "Blair Witch Project" [10] quality of first-person video? Tod's outright rejection of the disorienting high quality, low frame-rate video feed was evidence enough that that encoding was no longer a viable candidate, allowing us remove it from consideration in our lab study, and focus on this potentially stronger result.

We were interested in uncovering people's subjective reaction to different encodings of first-person-video. Very simply, *Do you like it or not?* We wished to divorce the content and any perceived task from the judgments. It is easy to conceive of tasks that make any of the encodings succeed or fail, so a task-oriented evaluation would reveal nothing. Instead, we had to impress upon the subjects that it was the quality of the video that they were judging, and assure them that it was okay for the judgment to be purely subjective and even instinctive. The scenarios that were viewed and the questions that were asked were designed to achieve this.

6.1. Experiment Setup

We recorded three 2-3 minute first-person video segments using a camera setup similar to the one described in the previous section. The first was a video of a trip through the grocery store (representing a crowded environment), the second was video of someone making breakfast for the kids (representing an indoor home environment), and the third was video of someone taking out the garbage (representing an outdoor scene). The goal was to make the camera motion and activities as natural as possible.

The three videos were then encoded in four different ways. *encFast (eF)* was sampled at 1fps and run through our visualization system. *encSlow (eS)* was similar, but sampled at .67fps. *encIdeal (eI)* was the "ideal" version, encoded at roughly 11fps (the fastest our camera system could record raw video frames) with an infinite bitrate budget. And *encChoppy (eC)* was encoded at 5fps at a comparable bitrate to the corresponding *encFast*.

The subjects were asked to watch all of the video clips in whatever order they desired, and were encouraged to do side-by-side comparisons. They had complete playback control (pause, rewind, etc.). The following questions were given to the subjects prior to the start of the experiment, and answers were solicited throughout.

What is your gut reaction? Rank the video feeds in order of preference. Describe the characteristics of each of the video clips. Why do you like it? Why don't you like it? If it was your job to watch one of these clips all day long, and there was no specific task involved, which would you choose? Why? Do any of these clips cause you physical discomfort? Which ones? Do any of the clips create confusion? If so, is it temporary or perpetual? Discounting the content, how do each of the clips make you feel? Have your preferences changed?

These questions were designed primarily to encourage the subjects to think critically about each of the clips. Obtaining a carefully considered ranking of the clips was our main goal. However, the responses would also help shed light on the underlying reasoning.

#	S	A	G	Initial Pref	Final Pref
1	M	20	T	eI, **eF**, **eS**, eC	eI, **eF**, **eS**, eC
2	F	60	F	eI, **eS**, **eF**, eC	eI, **eS**, **eF**, eC
3	M	40	F	eI, **eF**, **eS**, eC	**eS**, **eF**, eI, eC
4	F	40	F	**eF**, **eS**, eI, eC	**eF**, **eS**, eI, eC
5	F	60	F	eI, **eS**, **eF**, eC	eI, **eS**, **eF**, eC
6	M	30	T	eI, eC, **eF**, **eS**	eI, **eF**, eC, **eS**
7	F	30	T	eI, eC, **eF**, **eS**	eI, **eF**, **eS**, eC
8	M	30	F	eI, **eS**, eC, **eF**	eI, **eS**, eC, **eF**
9	M	20	F	**eS**, eI, **eF**, eC	**eS**, eI, **eF**, eC
10	M	30	T	eI, eC, **eF**, **eS**	eI, eC, **eF**, **eS**
11	F	30	F	**eS**, **eF**, eI, eC	**eS**, **eF**, eI, eC
12	M	30	T	eI, **eF**, eC, **eS**	eI, **eF**, **eS**, eC
13	M	20	T	eC, **eF**, **eS**, eI	eC, **eF**, **eS**, eI
14	M	60	F	eI, **eF**, **eS**, eC	eI, **eF**, **eS**, eC

Table 1. *Summary of results. # is the subject number,* S *is the sex,* A *is the age, and* G *indicates if the subject had any 1st-person-shooter game experience.* Initial Pref *is the gut reaction ranking given to each of the encodings, and* Final Pref *is the final ranking. References to our encodings appear in bold.*

6.2. Results

The following summarizes the data found in Table 1. 14 subjects participated in this study, 10 male, and 4 female, ranging in age from 20 to 60. All but two of the subjects preferred at least one of our encodings to the choppy encoding, and 4 of the subjects actually preferred our encodings to the ideal encoding that was used as a control. 6 of the subjects preferred eS to eF, and in all of these cases the preference was very strong. None of these 6 subjects had first-person-shooter game experience. 4 of the subjects changed their ranking of the encodings midway through the experiment, and in all cases our encodings were ranked higher.

6.3. Analysis

Our encodings faired much better than expected. Not only did 4 of the subjects rank them higher than eI, there were also 4 others who were explicitly on the fence, and saw definite benefits to our encodings. Our encodings also seemed to grow on people. 4 of the subjects changed their rankings towards the end of the experiment, moving our encodings higher in preference. Everyone in the study liked our encodings, regardless of how they ranked them. The following is a sampling of the positive qualities voiced by our subjects: *calm, smooth, slow-motion, sharp, artistic, soft, not-so-dizzy*. There were of course some negative characterizations, too: *herkey-jerkey, artificial, makes me feel detached, insecure*.

The clearest pattern was the subjects' dislike of eC. We will discuss the two exceptions to this a little later. Most stopped paying attention to eC early in the experiment because the quality, to them, was obviously much poorer.

Many of the subjects had a strong personal criterion that they used for judging the videos. For some, it was clarity of the images and for others it was the lack of choppiness. There were also those who were most influenced by nausea.

The subjects in the clarity camp (subjects 2, 5, and 14) were interesting because despite the clarity of the images in eF and eS, they were bothered by the momentary blurriness during the transitions. As a result, they preferred eI even though some of them indicated that the speed of the video was too fast. The clarity camp may have responded better to transitions that do not do alpha-blends.

The slightly longer interval between frames in eS made all of the difference for some. Subject 8 actually ranked eF the lowest because it was just too "herky-jerky". Subject 9 liked eS the best, but ranked eF below eI. eF "had a jolting, motion sickness feel." Others, on the other hand, had strong negative reactions to eS, because it was too slow and boring. There appears to be a strong correlation between an individual's lack of first-person-shooter game experience and their preference for eS. None of the subjects who preferred eS had any game experience. First-person-video is not something that people get a lot of experience watching, unless they play first-person-shooter games. With more experience, people may actually prefer the speed of eF.

This study helped explain why first-person-video can be so difficult to watch. It comes down to control and expectation. We are not bothered by our own first-person view of the world because we are controlling where we look. We can anticipate what the motion is going to feel like, and we know what to expect when the motion stops. When watching something through another person's eyes, however, that expectation is lost, so we are always playing catch-up. Subject 4 preferred our encodings over eI precisely for this reason. She said that eI was moving so fast that she could not pick up any of the details. Just as she was about to focus on the current scene to comprehend it, the view moved to something else. She liked that eF gave her the extra time to actually absorb what was going on.

Subjects 10 and 13 were the only ones who preferred eC over our encodings. Their reasons were quite different so we will consider them independently. Subject 10 simply preferred traditional video to our encodings. He could see the value in our encodings, and was not confused by them, but he felt detached watching them.

Subject 13 is an interesting outlier. Not only did he rank eC the highest, but he ranked eI the lowest! In a post-experiment interview we learned that he preferred the artistic quality of eC. It was edgy. He was bored by eI and found it a little bit nauseating. He also liked the artistic feel of our encodings, but ultimately the "predator" feel of eC is what drew him towards that one. Clearly, there is no one solution that would appeal to everyone.

6.4. Secondary Study

A secondary goal of the lab study was to determine if the morphing performed during transitions was helpful. As described in section 3, morphing can create better alignments between the images by making all matching points overlap exactly throughout a transition. It was not clear to us that the morphing was providing much benefit, and when the vision algorithm occasionally returned incorrect matching points, the morph looked startlingly bad.

The surprising result was that none of the subjects could discern any difference between *morphed* and *non-morphed* video clips when played side-by-side. It was only when the video was slowed down by a factor of eight that some of the subjects noticed a difference, although even then they only expressed a vague preference for one over the other. Stop-motion convinced all of the subjects that the alignment between images was indeed better in the *morphed* version.

We hypothesize two explanations for this result. (1) Our brains are so good at committing closure that unless there is perfect alignment between images, varying degrees of misalignment (to a point) are perceived as being the same. There are times when closure is being performed consciously, but for the most part this is a process that happens unconsciously, and people are only vaguely aware of it happening. (2) The dynamic content is what is interesting in a scene – the very content that does not get morphed because matching points cannot be found.

7. Conclusion

We have presented a visualization technique for displaying low-bit-rate first-person video that maintains the benefits of high resolution, while minimizing the problems typically associated with low frame rates. The visualization is achieved by performing a dynamic visual interpolation between frames using meta data captured from a digital pan/tilt compass or inferred using vision techniques. We have demonstrated with a field study that this technique is appropriate in a command center, in contrast with traditional low-bitrate encodings which may cause disorientation and physical discomfort. Our lab study confirmed that our visualization has wider appeal and may have application in many other contexts. 12 of our 14 subjects preferred our visualization to the current state-of-the-art given comparable bitrate budgets. The surprising result is that 4 of the subjects actually preferred our visualization to the high framerate, high quality video that was used as a control. These results are based on subjective preferences across three different domains, and are thus untainted by task-specific evaluations that would limit the generality of our findings.

8. Acknowledgments

Special thanks to Tod Ferguson and the UCSD Hazmat team. This work was supported in part by contract N01-LM-3-3511 from the National Library of Medicine and a hardware gift from Microsoft Research.

References

[1] International Organisation for Standardisation: ISO/IEC JTC1/SC29/WG11MPEG 98/N2457. 1998.

[2] M. Arisoylu. 802.11 wireless infrastructure to enhance medical response to disasters. In *Proc. AMIA Fall Symp*, 2005.

[3] P. Baudisch, D. Tan, D. Steedly, E. Rudolph, M. Uyttendaele, C. Pal, and R. Szeliski. Panoramic viewfinder: providing a real-time preview to help users avoid flaws in panoramic pictures. In *OZCHI '05: Proceedings of the 19th conference of CHISIG of Australia on CHI*, pages 1–10, Narrabundah, Australia, Australia, 2005. CHISIG of Australia.

[4] M. Brown and D. G. Lowe. Recognising panoramas. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1218, Washington, DC, USA, 2003. IEEE Computer Society.

[5] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. Efficient representations of video sequences and their applications. *Signal Processing : Image Communication*, pages 327–351, 1996.

[6] O. Komogortsev and J. I. Khan. Predictive perceptual compression for real time video communication. In *ACM Multimedia*, pages 220–227, 2004.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[8] S. McCloud. *Understanding comics: The invisble art.* Harper Collins Publishers, New York, 1993.

[9] N. J. McCurdy and W. G. Griswold. A systems architecture for ubiquitous video. In *Mobisys 2005: Proceedings of the Third International Conference on Mobile Systems, Applications, and Services*, pages 1–14. Usenix, 2005.

[10] D. Myrick and E. Sanchez. Motion picture: Blair witch project. 1999.

[11] S. M. Seitz and C. R. Dyer. View morphing. In *Proc. SIGGRAPH 96*, pages 21–30, 1996.

[12] D. Steedly, C. Pal, and R. Szeliski. Efficiently registering video into panoramic mosaics. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 2*, pages 1300–1307, Washington, DC, USA, 2005. IEEE Computer Society.

[13] R. Szeliski. Image mosaicing for tele-reality applications. In *WACV94*, pages 44–53, 1994.

[14] M. Weiser. The computer for the 21st century. *Human-computer interaction: toward the year 2000*, pages 933–940, 1995.

[15] 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.