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

Neil J. McCurdy[1,2], William G. Griswold[1,2], and Leslie A. Lenert[2,3]

[1] Department of Computer Science and Engineering
[2] California Institute for Telecommunications and Information Technology
University of California, San Diego, La Jolla, CA
[3] 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

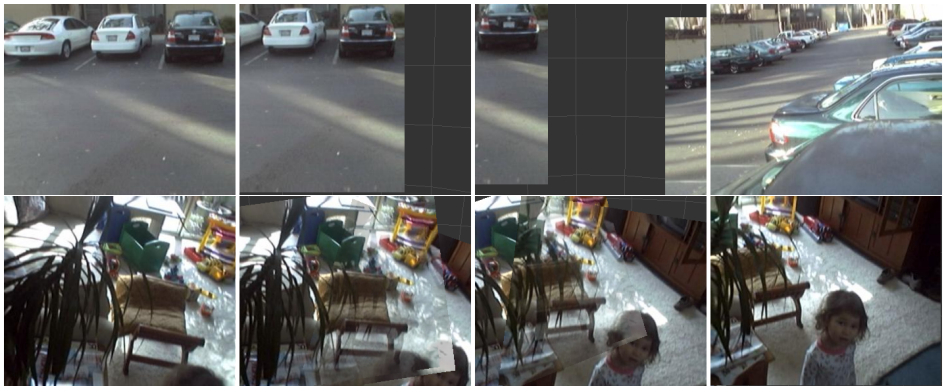

Fig. 1. *Snapshots of two transitions in progress. The top row depicts a camera pan from left to right where the frames do not overlap. The bottom is a morph from the frame on the left to the frame on the right as the camera pans down and to the right to look at the child. The live experience is one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

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. Even in a relatively unchanging outdoor environment where there is a large field of view, a viewer can become disoriented looking at the camera's view of the ground when the camera operator looks down to avoid obstacles.

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 subsequent 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 [1]. 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 some short video clips of transitions downloadable from the web (`http://ubivideos.homeip.net`).

No knowledge of the camera's position is required, unlike the previous work involving inter-camera transitions [1]. The assumption is that the amount of positional change in the interval between two frames is not significant, and the results of our user studies confirms this. Even without the explicit representation of position, however, the viewers still have a sense of movement through the environment. Not only is there the illusion of movement similar to the illusion experienced when watching any sequence of frames, but there is real movement as well. The manner in which we align the subsequent frames when there is frame overlap, and the transition between the frames, creates the sensation of movement. At times the alignment will cause the entering frame to start off smaller than it really is, and then grow in size (zoom in) until it fills the screen. This zooming creates the appropriate sensation of moving forward (or conversely, backward) through the environment.

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 [2].

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 in detail, and section 4 discusses related work. We present our field and lab studies in sections 5 and 6, and then conclude.

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 hap-

pening 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.

On a more personal level, our user study revealed that people may be interested in viewing live first-person streams of their distant loved-ones. Grandparents, for example, may want to join the grand kids on a trip to the zoo.

2.1 Hazardous Materials Disposal

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 interviewed the Hazardous Materials Business Plan Manager (hereafter referred to as Tod) at the University of California, San Diego (UCSD) to determine how he thinks a live video feed could be used in managing a hazmat scene.

Although Tod's team is trained and ready to handle disaster scenarios, a typical incident is thankfully fairly mundane. On a university campus, the most typical emergency condition is a fire in a laboratory that contains toxic chemicals. After the fire has been put out and the situation has been stabilized there is often a substantial cleanup effort that can take anywhere from days to weeks – all of which must be performed in full suits with masks and respirators. Tod's primary concern during one of these responses is ensuring the health and safety of his team. As a commander, it is his 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. Currently, Tod does not operate with any visuals of the scene, and because of this, he often rushes to an incident so that *he* can be the first person to enter the environment. This way he can use his memory of the conditions to help him make future judgments.

As one of the most experienced members on his team, Tod asserted that he could use the information he receives from video feeds to help identify hazards, to coach the less experienced members who are on the inside, and most importantly to see that his team is active and healthy. Fatigue and heat exhaustion is a serious problem in this line of work, and the hero mentality that is common among first responders often causes these symptoms to go unreported. Tod said he would go so far as to make his team members hold their air gauges to the camera since he does not necessarily trust their self-reports – not because the team members are untrustworthy, but because the symptoms of fatigue, the conditions of the environment, and the cumbersome suits that are being worn could cause errors in the reading of the displays.

Our first reaction to the requirement for health and safety readings was that sensors that report such things as heart rate, body temperature, etc. could transmit this data at a much lower bandwidth. Tod, however, was eager to invest in video because the same information could be conveyed by observing the body posture and the activity of his team, as well as support the other functions cited above. He said that feeling like he is actually there, in the environment, is important to him.

There are other benefits to having video. For example, what should be a simple task, like finding a shut-off valve, becomes difficult when the people who know the

environment (lab managers, for example) cannot see what the people on the inside are seeing. Not only would the lab manager be operating from memory, but the memory is likely outdated since the conditions in the lab may now be very different.

The preference for head-mounted cameras over pan-tilt-zoom tripod-based cameras is motivated by shortcomings of fixed cameras. There is no reason why fixed cameras cannot also be supported, but the dynamic nature of a hazmat response suggests the need for mobile cameras. It would be very difficult to position fixed cameras, even with pan, tilt, and zoom capabilities in a place that would give enough coverage of the environment while still providing detail of the activities of the team members. Multiple cameras would be required if there are multiple work environments with obstructions (e.g., walls, furniture, or debris) between them. A head-mounted camera has the benefit that it is almost always focused on something that is of importance to at least someone (the camera operator).

2.2 Networking Challenges

Operating in real-world environments is always challenging, but we are continually been surprised by just how challenging the conditions are for wireless networking. In prior work, we attempted to stream video during county-sponsored disaster drills across an 802.11b wireless mesh network [3]. We were confident of success, given that we were using a network that was carried on scene, self-configures, and is battery powered. Two drills later, in which thousands of people were involved, with helicopters, fire trucks, and the media interfering with the network in numerous ways, we are still learning how to deal with the realities of wireless networking in disaster response. And being just drills, the conditions of a real disaster scene would be different still. The following list of network challenges is derived from our experience operating in these environments.

Weak infrastructure support. One cannot rely on an existing network to exist, with the possible exception of cellular networks, which have cell towers far removed from the incident. In large scale disasters, though, cellular networks have been overloaded, rendering them them useless for extended periods [4].

Unreliable networking. For networks brought on site, expect frequent network congestion and failures, causing device disconnects. Interference is caused by both natural phenomena and competition with other networks deployed in the same space (pre-existing or imported for the response).

Low bitrate. Even on an 802.11b wireless network which has an effective single-source bitrate of 6.2Mbps, with three cameras on a wireless mesh, we have not been able to rely on much more than 100kbps per stream. Noise in a real-world environment contributes to this loss of throughput, as does the fact that the total throughput drops drastically as more nodes are added to the system. Empirical studies have shown that with more than eight nodes, total throughput decreases to roughly 2Mbps [5]. Also, since each camera is mobile, the perceived signal strength of the local access point (and thus the bandwidth of the connection) will vary depending on the location of the camera and intervening obstructions.

2.3 Video Compression Challenges

The conditions that have been outlined so far present a significant challenge for video compression. Let us first motivate the need for compression. Uncompressed CIF (352x288 pixels in RGB24 format) video playing at a typical 30fps requires a 69Mbps pipe. With spatial compression, each frame can be reasonably compressed from 297KB to a 12KB JPEG image. This compression technique, called MJPEG (motion JPEG) reduces the bandwidth requirements to 2.8Mbps. Temporal (inter-frame) compression like that provided by MPEG makes it possible to reduce the bitrate to the 300kbps range without sacrificing much in terms of quality.

At the most basic level of detail, an MPEG-style encoder works roughly as follows: A group of pictures (GOP) begins with an I-frame (which can be thought of as a JPEG encoded image), and is followed by multiple P-frames (a predicted frame which encodes the difference between the current frame and the previous frame). A P-frame is more than just the simple difference between frames, though – the motion of objects between the frames is taken into account and is encoded as motion vectors. We will ignore B-frames since they are not important for this discussion. The length of a GOP is usually specified as a parameter to the codec, but may also be determined by scene changes. None of the P-frames are useful if any of the previous frames are lost, so I-frames are important for error recovery in streaming scenarios, or to facilitate random access playback for locally stored media.

If there is significant redundancy between frames, the difference between P-frames will be small, and a high level of compression will be possible. If, on the other hand, there is a lot of rapid panning (a common characteristic of first-person video), the differences between frames will be great, and the P-frames may offer no better compression than the I-frames. With traditional video codecs, there is no way around this. Without temporal redundancy, there is little chance of doing better than MJPEG.

There are three options for rate-limited video that does not have temporal redundancy (assuming the use of traditional codecs). (1) The frame rate can be reduced, (2) the encoding quality of each frame can be reduced, or (3) a combination of these two alternatives can be used. Reducing the frame-rate makes the video choppy and jittery – a condition that many users in our user study could not tolerate. Reducing the frame-rate below 5fps changes the experience of video to a sequence of still images. As the frame rate drops, it becomes difficult to track objects, and eventually it is even difficult to orient yourself in the scene. The other option, reducing the image quality, has similar problems. While the motion will be smooth, the blurriness of the image may make it difficult to identify objects.

To get the bitrate within 100kbps (to support three streams on our mesh network), we have to sacrifice both the frame rate and the image quality in order to have the feed continue to look like video (in other words, stay above 5fps). We found, and our user studies confirmed, that video at this quality is really not acceptable in most circumstances. The choppiness and blurriness induce nausea and headaches.

The remaining choice, then, is to drop below 5 fps and optimize instead on image quality. Referring back to our MJPEG calculation above, the average JPEG encoding size of 12KB at 1fps translates into a bitrate of 96kbps. Dropping down to one frame per 1.5 seconds comfortably keeps us under 100kbps and even gets us under the 64kbps 1xEVDO cellular network limit.

In an unreliable network, MPEG has an additional problem: packet-loss results in a garbled image until the next I-frame is received. If temporal rendundancy is low,

anyway, it may be best to limit the effects of packet-loss to individual frames by using an MJPEG-style compression scheme instead. This is our motivation for using MJPEG.

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 [2]. 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. An interesting result of our lab study is that almost no users were able to discern the difference between segues that involved roughly aligned frames and those that were more accurately aligned. In fact, of the few users that could discern a difference, some of them actually preferred the rough registrations. Closure is that powerful.

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 [6] 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.

For each incoming frame, our rendering engine looks at the list of matching points and determines if there is a match with the previous frame. To help remove erroneous matches that made it through the RANSAC filter [6], we further filter the data based on the expected usage pattern of the camera. For example, since the camera is mounted on a person's head, we may be able to assume that if matching points correspond to a side-to-side tilt of the camera by more than 45° the matching points are erroneous.

The 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 control 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, with the help of closure this technique can 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 [1]. It requires knowledge of the positions of the cameras, as well as the orientations, effectively 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 [7]. 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. Using 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. Irani reports significant compression improvements over MPEG which was the standard in 1996, and even using today's standards the quality achieved at 32kbps is impressive. Unfortunately, mosaic-based compression cannot be used for encoding mobile first-person video because the 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 [8] (also known has MPEG4-10) represents the current state-of-the-art. MPEG4-2 [9] (commonly referred to simply as MPEG4), the previous state-of-the-art, is more established and is more likely to be supported in media players. There is little perceptible difference between codecs that support these standards when compressing first-person video at low bit-rates. This is not surprising considering the absence of opportunities for temporal compression.

A non-traditional approach to video compression proposed by Komogortsev, varies the quality of the video based on where the viewer is looking [10]. 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 the environments we support because the network latency is high (4-5 seconds); the gaze direction would have to predicted far in advance.

There has been substantial work on generating panoramas from still photographs [11, 12]. Real-time dynamic creation of panaoramas on a handheld camera device has been used to help with the creation of a static panorama [13]. Panoramas can also be efficiently created from movie cameras assuming the camera's position is relatively static [14]. All of these techniques require some way to match points between images. We rely heavily on Lowe's SIFT algorithm [6], specifically the Autopano implementation of it (`http://http://autopano.kolor.com/`).

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 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 bulkers also learn how to handle hazardous materials and obtain first-hand experience with the properties of the chemicals with which they are dealing. For example, it is not uncommon for labs to mislabel their materials, which can result in a dangerous chemical reaction when the materials are combined in the drums.

The bulking process typically takes one to two hours depending on the amount of waste material (roughly 250 gallons on average). During this time the bulkers are isolated in a closed room because they are the only ones wearing equipment to protect them from the noxious fumes. Tod, the team leader introduced in section 2, often worries about the health of his team during these exercises, and looks forward to using a video system similar to the one tested so that he can check on the bulkers periodically. When we suggested that a permanent, wired camera might be more appropriate for this particular situation, he re-emphasized how important preparation and training were in his field. He wants his team to be training in the actual equipment that will be worn during emergencies. They need to feel comfortable using and wearing it, and Tod needs to have enough experience with it to trust it.

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. It was also clear after our first interview with Tod that the bulkers were not going to tolerate any setup that would impede their work. This is a dirty, tiring job, and they were going to have little patience for anything that made them stay suited up for longer than normal.

We attached a disassembled Logitech webcam ($\sim$\$100) to the front of the mask, and sewed a tilt sensor manufactured by AOSI ($\sim$\$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 ($\sim$\$2000) which was placed in a small backpack. Consistent with Tod's dictum that his bulkers work with the same gear as in incident response, 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 Vaio lacked the PC Card slot we needed for a 1xEVDO modem, however, so we connected via 802.11 to one of our wireless mesh network nodes, and had the traffic routed through

1xEVDO from there. 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. Tod's task was to 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. We detected what looked like severe congestion on the 1xEVDO link at the 50 minute mark. All of the results reported in this section discount the first 6 minutes of data (setup time), and the last 14 minutes of congested data.

The bulking task turned out to be very demanding for our visualization system. Bulking is not only labor intensive with constant activity and motion, but is also conducted in close quarters (a roughly 3-5 foot zone), leaving little opportunity for the viewer of the video to get the perspective that comes with a wider field of view. The camera motion was also unlike anything we had seen before in artificial drills. The bulker was constantly bending down to his left to pick up a drum (weighing up to 27KG), and then placing it in the sink. This caused the view to move back and forth between what we will call the origin (0° longitude, 0° latitude) and -90° longitude, -90° latitude. These are quite extreme movements given the limited field of view of our camera (44° long, 33° lat). The bulker indicated that he barely noticed our equipment, so we judge that these extreme movements are representative for this task. It is likely that similar motion models would be found in the cleanup phase of a burned-out laboratory, where the task resembles a demolition effort.

Tod reacted favorably to the visualization. The lack of reaction is probably most telling, considering the novelty of the visualization for him. He paused for a second as he absorbed what he was seeing, and then began: "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." Tod then began describing the bulking process, at first just giving background information that did not require access to the visuals. After this brief interlude, I asked him if he could tell what was going on. "Yeah, I can tell that Sam is doing the bulking... This gives me a good look at the funnel area so that we would see reactions if there was a chemical reaction. Normally that's displayed through vaporization. If you're lucky here (pointing to the screen) you might get a little bit of that. 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... This definitely lets me know that they're still working. That's really the important part."

Tod expressed an interest in flipping through the still photographs so that we could really study individual 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. Tod was also intrigued by the time-stamping and indexing of all the images, as reconstructing timelines of an event for post mortems is currently difficult because time pressures and distractions undermine recordkeeping.

Tod then noticed the birdseye view that uses arrows and cones on a black backround to represent the orientations of the camera views that are currently active. This helped him orient, and he started using this view to illustrate where things were and what the other bulker might be doing.

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 like this here (pointing to the birdseye view). This is telling me the orientation in the room. I like these two. 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."

Throughout the exercise, Tod weighed in on why our visualization tool would be effective in a control center environment. "...to have a visualization that adds credibility to your discussion, 'this is what we were doing, and this is the size of the equipment', and evaluating what resources are going to be needed for subsequent entries, and hopefully we can get this from that single entry. You can't get past the benefits of the visual. It brings us to a whole other level of safety, assurance, and competence."

He went on to explain how decisions are made by gut feelings, based on the skills of the people that are involved and on his comfort level with those people. It is all about contact, he said, and anything that increases contact is going to improve these decisions. Contact is especially important when the lives of people you deployed are at stake. "Video gives you a better gut feeling to what is going on. Video gives you another form of contact. It builds trust in your decision making."

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. It really is the combination of the fact – it's another piece. It's not the all-empowering 'this is the tool', you know. You don't want that, because if it didn't work you don't want to all of a sudden – 'oh we can't do anything because it's not working' You don't want that. What you want is good components that can go in and help add, and help make better decisions... It's appropriate. It's not overwhelming. It doesn't seem to be large, cumbersome, overly difficult."

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: "I can see where I would put in some more wide angles. I assume I could take one of these cameras and

just set it [in the environment]... Most of the events are quick and dirty. You wouldn't go with a stationary camera unless it was a prolonged cleanup."

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 encoded at bitrates comparable to the original experiment, using FFMPEG's MPEG4 codec, which was as good as the experimental H264 codec described earlier (`http://ffmpeg.sourceforge.net`). His reaction surprised us: "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 [15]. "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."

We hypothesize that this intermediate frame rate overtaxed Tod's closure capabilities. At 30fps, the motion between frames is small enough for the result to appear smooth. At 0.5fps, with high-frame-rate segues, there is both smoothness and ample time to dwell on each target frame. At 5fps, there is little time to take in any individual frame, and there is too much happening between frames, too quickly.

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. We were now curious if our visualization method would have broader appeal. Might it actually be an alternative to the sometimes nauseating, "Blair Witch Project" [16] quality of first-person video? Would people choose to watch first-person live video feeds of distant loved ones if given the opportunity? Would Grandma want to virtually join the grandkids on a trip to the zoo?

We were interested in uncovering people's subjective reaction to different encodings of first-person-video. Very simply, *Do you like it or not?* This means that we had to somehow divorce the content and any perceived task from the judgments. Obviously, if the goal of watching the video was to read the text of a poster on a distant wall, for example, then image clarity would be the most important quality. Likewise, if the goal was to detect whether or not a big red ball bounced through the scene, the frame rate would be most important. Our task, then, was 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, but with a baseball cap replacing the hazmat mask. *Groceries* was a video of a trip through the grocery store (representing a crowded environment), *Breakfast* was video of someone making breakfast for the kids (representing an indoor home environment), and *Garbage* was video of someone taking out the garbage (representing an outdoor scene). Each of these videos was designed to record a task to make the camera motion and the 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*.

Starting with *Groceries*, the subjects were asked to begin watching each of the encodings sequentially, but were then encouraged to watch them all in parallel so that they could do side-by-side comparisons. The subjects were allowed to resize the video windows, and could pause, rewind, and fast forward through the clips as desired. *Breakfast* was viewed next, followed by *Garbage*. The following questions were given to the subjects prior to the start of the experiment, and answers were solicited throughout. The subjects were encouraged to alter their answers if subsequent clips revealed something new.

- 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?
- Would you enjoy watching any of these clips (assuming interesting/relevant content)? For example, to see kids, grandkids, friends, etc.
- 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? If so, what is the new ranking?

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, we also wanted to analyze the responses to help shed light on their underlying reasoning.

6.2 Hypotheses

We expected the encodings to be ranked in the following order of preference: (1) eI, (2) eF, (3) eS, and (4) eC, but thought some may prefer eF or eS over eI. eS was a last-minute addition to the study after one of the authors felt a little queasy while watching eF. We hypothesized that dropping the frame rate a little may make the difference for some subjects prone to motion sickness. eS also encodes close to 1xEVDO network bitrates.

6.3 Results

Subject	Sex	Age	Game Exp	Initial Pref	Final Pref	Nauseating
1	M	20	T	eI, **eF**, **eS**, eC	eI, **eF**, **eS**, eC	eI, eC
2	F	60	F	eI, **eS**, **eF**, eC	eI, **eS**, **eF**, eC	eC
3	M	40	F	eI, **eF**, **eS**, eC	**eS**, **eF**, eI, eC	eI
4	F	40	F	**eF**, **eS**, eI, eC	**eF**, **eS**, eI, eC	eI
5	F	60	F	eI, **eS**, **eF**, eC	eI, **eS**, **eF**, eC	**eF**, eC
6	M	30	T	eI, eC, **eF**, **eS**	eI, **eF**, eC, **eS**	eC
7	F	30	T	eI, eC, **eF**, **eS**	eI, **eF**, **eS**, eC	eI, eC
8	M	30	F	eI, **eS**, eC, **eF**	eI, **eS**, eC, **eF**	eI
9	M	20	F	**eS**, eI, **eF**, eC	**eS**, eI, **eF**, eC	eI
10	M	30	T	eI, eC, **eF**, **eS**	eI, eC, **eF**, **eS**	**eF**
11	F	30	F	**eS**, **eF**, eI, eC	**eS**, **eF**, eI, eC	eC
12	M	30	T	eI, **eF**, eC, **eS**	eI, **eF**, **eS**, eC	eI, eC
13	M	20	T	eC, **eF**, **eS**, eI	eC, **eF**, **eS**, eI	eI
14	M	60	F	eI, **eF**, **eS**, eC	eI, **eF**, **eS**, eC	eC

Table 1. *Summary of results.* Game Exp *stands for 1st-person-shooter game experience.* Inital 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.*

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 visualizations to the choppy encoding, and 4 of the subjects actually preferred our visualizations to the ideal encoding that was used as a control. All of the subjects reported that some of the video clips caused some physical discomfort (nausea, mostly). eI and eC were the common culprits for this, but two individuals had trouble with eF. 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 visualizations were ranked higher.

6.4 Analysis

We were surprised by how well our visualizations were received. Not only did 4 of the subjects rank our visualizations higher than eI, there were also 4 others who were explicitly on the fence, and saw definite benefits to the visualizations. Our visualizations also seemed to grow on people. 4 of the subjects changed their rankings towards the end of the experiment, moving our visualizations higher in preference. Everyone in the study liked the visualizations, 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. This lack of consideration may explain the occasional absence of eC but the presence of eI in the *Nausea* column in Table 1.

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 clarity camp (subjects 2, 5, and 14) is interesting because it was not until the end of the study that we realized what bothered them about our visualizations. Subject 5 kept reiterating that the characteristics she sought were "slow and clear", and yet she chose eI over eS. The image quality of eI and eS should have been identical, and eS did not have the fast, jittery quality that the "slow" request was an obvious reaction to. Subject 14's similar responses solved the puzzle. Although the image clarity of the individual images is high in our visualizations, the alpha-blend performed during transitions causes a temporary blurriness since the alignment between images is not perfect. Transitions that do not use an alpha-blend have a certain appeal, but we ultimately chose to include the alpha-blend in the clips used for this study because, in our opinion, the alpha-blend makes the transitions feel smoother and calmer, as well as assisting in closure. It would be interesting to get the clarity camp's reaction to non-alpha-blended transitions. For those who used the lack of choppiness as their main criterion, the non-alpha-blended transitions would probably be unfavorably received.

It was fortunate that we added eS to the study, because 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. We surmise that with more experience, people may actually prefer the speed of eF.

This study helped us understand why first-person-video can be so difficult to watch. It mostly boils down to control and expectation. Obviously we all have experience watching our own first-person-video every day of our lives. Why are we not bothered by it? We are controlling where we look, and because we are controlling it, we anticipate what the motion is going to feel like, and we have a pretty good idea of 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 visualizations 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 visualizations. Their reasons were quite different so we will consider them independently. Subject 10 simply preferred traditional video to the visualizations. He could see the value in the visualizations, 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 visualizations, but ultimately the "predator" feel of

eC is what drew him towards that one. Clearly, people are different – there is no way we could ever create a solution that appeals to everyone

The different scenes did not appear to make any difference to the subjects' preferences. None of the scenes were responsible for a change in ranking. People seemed to enjoy watching the *Breakfast* video the most because of the presence of the kids. This video was probably responsible for all but one of the subjects indicating that they could see themselves enjoying watching live first-person video of their loved ones. In this context, some of the subjects who liked eI the best thought our visualizations would be more appropriate. This can be attributed to the fact that many found our visualizations easy to watch. Details are probably not very important in this context, so the lower frame-rate would not be a factor.

6.5 Secondary Study

During this lab study we took the opportunity to investigate the subjective value of the morphing performed when transitioning between frames that were aligned via point matching, as described in section 3. It was not clear that morphing was providing much benefit, and when the vision algorithm occasionally returned incorrect matching points, the morph looked startlingly bad.

We had our subjects do side-by-side comparisons of a morphed and non-morphed version of the *Garbage* video encoded as eF. They also made a similar comparison with the *Groceries* video, although this time the transitions were slowed down by a factor of 8. None of our subjects could discern any difference between the morphed and non-morphed versions of the *Garbage* video. After watching the *Groceries* video, most of the subjects still barely noticed a difference, but many had a vague preference for one over the other. These preferences are not surprising in light of the range of preferences cited above: 4 preferred the non-morphed version because it was softer and rocked less, and 4 the morphed version because it was sharper.

All of the subjects were able to see the differences once they were pointed out, and stop-motion revealed that the alignment between the morphed images was much better. So why is it that the subjects had such a difficult time seeing the differences themselves? We hypothesize two explanations. (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 interesting content of the scene is the dynamic elements – the very content that does not get morphed because no matching points are found on them between frames.

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 showed that people may actually choose to watch such video for entertainment since it has the unexpected benefit of eliminating the "Blaire Witch Project" [16] effect – the nausea-inducing jumpiness typical of first-person video. Indeed, 4 out of 14 subjects in our study actually preferred this visualization to the high frame-rate, high quality video that was used as a control.

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. McCurdy, N.J., Griswold, W.G.: A systems architecture for ubiquitous video. In: Mobisys 2005: Proceedings of the Third International Conference on Mobile Systems, Applications, and Services, Usenix (2005) 1–14
2. McCloud, S.: Understanding comics: The invisble art. Harper Collins Publishers, New York (1993)
3. Arisoylu, M.: 802.11 wireless infrastructure to enhance medical response to disasters. In: Proc. AMIA Fall Symp. (2005)
4. McKinney, Company: Post 9-11 Report of the Fire Department of New York. (August 2002)
5. Vasan, A., Shankar, A.U.: An empirical characterization of instantaneous throughput in 802.11b wlans. `http://www.cs.umd.edu/~shankar/Papers/802-11b-profile-1.pdf` (2006)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2) (2004) 91–110
7. Irani, M., Anandan, P., Bergen, J., Kumar, R., Hsu, S.: Efficient representations of video sequences and their applications. Signal Processing : Image Communication (1996) 327–351
8. Wiegand, T., Sullivan, G.J., Bjntegaard, G., Luthra, A.: Overview of the h.264/avc video coding standard. IEEE Trans. Circuits Syst. Video Techn. **13**(7) (2003) 560–576
9. : International Organisation for Standardisation: ISO/IEC JTC1/SC29/WG11MPEG 98/N2457. (1998)
10. Komogortsev, O., Khan, J.I.: Predictive perceptual compression for real time video communication. In: ACM Multimedia. (2004) 220–227
11. Szeliski, R.: Image mosaicing for tele-reality applications. In: WACV94. (1994) 44–53
12. Brown, M., Lowe, D.G.: Recognising panoramas. In: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA, IEEE Computer Society (2003) 1218
13. Baudisch, P., Tan, D., Steedly, D., Rudolph, E., Uyttendaele, M., Pal, C., Szeliski, R.: 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, Narrabundah, Australia, Australia, CHISIG of Australia (2005) 1–10
14. Steedly, D., Pal, C., Szeliski, R.: Efficiently registering video into panoramic mosaics. In: ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 2, Washington, DC, USA, IEEE Computer Society (2005) 1300–1307
15. Weiser, M.: The computer for the 21st century. Human-computer interaction: toward the year 2000 (1995) 933–940
16. Myrick, D., Sanchez, E.: Motion picture: Blair witch project. (1999)