Layered Sensor Modalities for Improved Feature Detection

Dan Hestand
Computer Science Department
University of Massachusetts Lowell
One University Avenue, Olsen Hall
Lowell, Massachusetts 01854
Email: dhestand@acm.org

Holly A. Yanco
Computer Science Department
University of Massachusetts Lowell
One University Avenue, Olsen Hall
Lowell, Massachusetts 01854
Email: holly@cs.uml.edu

Abstract—Past attempts at using multiple sensor modalities in search and rescue activities have resulted in missed detection of victims and operator disorientation. This is largely due to information overload when the operator is presented with multiple displays of information requiring attention refocus. We propose a method wherein the multiple sensor inputs are layered into a single, integrated visual display. Such a display might eliminate most missed detections and alleviate operator disorientation by allowing the operator to focus on a single display with visual cues to aid in the use of sound and other sensor modalities. We have conducted initial trials with multiple sensors in the NIST Urban Rescue Test Arena to validate the concept and aid in the construction of appropriate display modes and operations.

I. INTRODUCTION

Current robot systems for urban search and rescue (USAR) typically send a video feed back to the operator. With a teleoperated robot, which most USAR robots currently are, the operator must use the video feed to navigate as well as to look for victims. With a great deal of operator overhead spent thinking about safe navigation, victims could be missed in the video feed, particularly since victims may be buried or covered in dust.

A study of the USAR systems at the AAAI Robot Competition in 2002 found that operator performance was impaired by the lack of information integration in the systems’ user interfaces [1]. For example, one region of the screen may contain the video feed, another region may have a window with some sort of sensor ranging data (sonar, ladar, etc.), status information about the robot is in another area of the screen, and yet another region may display a map of the area traveled so far. This requires a system’s operator to change his focus of attention in order to learn about different sensor readings. However, since the video feed is most commonly used for navigation, most of the attention is spent looking at the video screen. Creating an interface that overlays sensor information on top of the video image could help to provide a single focus for the operator, which should help him to avoid missing some critical information on another part of the screen.

Previous work has shown that combined imagery or the use of iconographic image representations have potential for improving the ability to detect features in the combined imagery that may not be obvious in the individual component images. For example, a thermal image and a camera image were combined in [2, 3]. The processing was on a single frame, yet shows the potential for this type of overlay.

The only competitor at AAAI-2002 who was using a FLIR camera to detect heat signatures of victims needed to flip the video window to the FLIR camera when the operator saw a potential victim. This required the operator to be able to see the victim in the color video screen in order to decide to flip the camera view over to the FLIR. However, this could cause the operator to miss victims. In fact, a user study at USF found that an operator of another system that needed to flip between video and FLIR did indeed miss seeing a victim while driving; an observer looking over the operator’s shoulder pointed out the victim in the video image [4, 5].

When looking for victims, using multiple sensor modalities can improve the chances of finding a trapped person. All systems entered into the AAAI competitions in 2002 and 2003 had a video camera to send data back to the operator. To deal with low light situations, some competitors attached flashlights to their robots with duct tape. Some systems used sonars for ranging; others using laser ranging. One competitor had a FLIR camera for heat detection. Another competitor put a wireless microphone on the robot so that the operator could listen to the environment around the robot. No competitor looked for CO₂ emissions, although some of the victims were rigged for this. No one system included a color video camera, FLIR camera and audio sensors.

In this work, we look for victims using a color video camera with IR LEDs automatically activated in low light, a FLIR camera, two microphones in order to calculate differential audio, and a CO₂ gas sensor. Rather than present four separate streams of sensor values in separate windows, we create a single, coherent view of the sensor readings by layering the
FLIR, sound and CO₂ data on top of the color video image.

II. METHODS

A. Sensor Selection

For hardware, we chose two color-CCD surveillance cameras with infrared LED low-light illumination for the visible imagery. Thermal imagery was obtained via a commercial-grade thermal imager. Audio data was provided by microphones built into the surveillance cameras. Finally, we used an industrial duct-mount CO₂ sensor to provide access to a gas channel.

The color cameras were mounted in a fixed position with a separation of 10 inches and a height of about 18 inches above ground level. The thermal imager was mounted as close as possible to the left-most camera to provide the closest alignment of the imager and camera apertures. The imagery from all three cameras was streamed into video cameras for storage onto tape and later processing. Since the microphones were built into the color cameras, we had no additional mounting requirements for them. However, since we wanted to be able to process the audio data separate from the video, we streamed the audio channels into a DAQ card for recording into .WAV files. The CO₂ sensor was mounted on the front of the platform at roughly 6 inches above the ground. The motivation for this configuration is that this would be the maximum height of a victim’s face lying on the ground. The data from the CO₂ sensor was streamed with the audio data into the DAQ card.

B. Data Source

The best source of baseline data with known sensor modality inputs is the NIST Urban Rescue Test Arenas. We made data collection runs during one day at the NIST site. The Test Arenas are three constructed areas consisting of partitions, furniture, simulated victims, and an array of environmental hazards [6, 7, 8]. We chose to perform data collection runs in the Yellow and Orange arenas. Simulated victims are constructed from display mannequins with heating pads wrapped around the torso to simulate body heat.

The Yellow arena is the most structured environment and has easily identifiable victims. The course is level and mostly uncluttered with debris, so data collection runs could be easily made. While the Yellow arena is the most structured, it does have sections with little or no illumination, confusing backgrounds (e.g., wall paper with closely spaced vertical stripes, mini-blinds, and mirrors at ground level), and areas where victims can be obscured. Audio input was provided by prerecorded voices on tape being played in a loop. There was no CO₂ input in the Yellow arena during our data collection runs.

The Orange arena is less structured than the Yellow arena but still largely traversable. There are corridors and rooms with debris strewn all around on the floor. The debris includes pipes, bits of wood, and large quantities of loose paper. There are also collapsed floors and areas that are only accessible by traversing a collapsed floor section. The victims are mostly obscured and in pinned or trapped positions so that access and identification is much more difficult than in the Yellow arena. Some of the mannequins in the Orange arena have a connection to a CO₂ source. The gas is emitted in a continuous stream instead of the small bursts as would be seen with real victims. For the purposes of validating the use of gas sensors, though, this is sufficient.

There is also a Red arena which presents the most unstructured environment. This arena has piles of rubble, collapsed and collapsible floor sections, and very difficult terrain features for navigation. We did not collect data from this arena for the purposes of this study since the sensor modalities present here were the same as in the Orange arena, only in a more difficult environment.

C. Data Fusion

Our approach to integrating the data is to avoid computationally expensive techniques. This notion suggests using techniques that are commonly available in both hardware and software. For example, with the video imagery we chose to use the standard video editing techniques of chroma matting with two video streams, one of which was made to be semi-transparent. Since the IR imagery had a well-defined background color (black) when there was no heat present, we chose this color to matte against. This technique is well-known and is often referred to as blue-screen or green-screen since these are the most common colors to matte with. In other environments or with other thermal imagers, the background color may not be black, but the chroma matte technique is flexible enough to allow selection of other colors and also selection of a range of colors about the chosen color.

For the audio, we chose to use the relative amplitude of the left and right microphones to provide a directional cue. To make this work we first have to eliminate any differences between the two microphones by normalizing the audio data. This relative amplitude is then converted into a linear position for an icon placed in the video imagery. So for a sound with a larger amplitude in the left microphone, the icon would be placed close to the left edge of the image. Since the audio data collection is at a higher rate than the video, we can easily update the icon placement with each frame of the video. So the operator should see an icon at the top of the screen sliding back and forth relative to a line representing the equal amplitude position giving a visual cue without requiring the operator to wear headphones. Since the audio channel data is available, if the operator requires, the data can be presented through headphones. This appears to be less useful, however, than simple visual cueing might be. For example, the rescue site may be noisy, the operator may need to be able to hear what others near him are saying, or the operator may be in a biohazard suit and unable to wear a system’s headset.

For the CO₂, we have a measurement ranging from 0 ppm to 2000 ppm. Since there is no directionality associated with the gas sensor, we have chosen to display the presence of gas by a sliding scale that increases with increasing gas concentration. Again, this is an iconic representation and is overlaid onto the video imagery frame-by-frame. If multiple gas sensors of
differences are not easily correctable and we have only been able to make adjustments for short segments of video where across a scene. Again, zooming of the two video streams a narrow opening that restricted our ability to enter without opening were not obvious or completely blocked. However, Overlaying the two images indicates that there may be a victim component visible and IR image data and the results of layering this structure by without a second glance, particularly if the values are 0.625 and 0.488, respectively.

There is additional magnification due to the differences in the optical characteristics of the lens systems used on the two different imagers. This difference shows up in magnification of the image and acceleration of the IR image when panning across a scene. Again, zooming of the two video streams can correct for the magnification. However, the slew rate differences are not easily correctable and we have only been able to make adjustments for short segments of video where the panning is minimal. Use of interchangeable lenses would help minimize this problem; we are investigating a FLIR camera with this ability.

III. RESULTS

Some results from our data collection in the Yellow Arena are shown in Figs. 1 and 2. These figures display the component visible and IR image data and the results of layering the two. Fig. 1 depicts a victim that is obvious in the visible imagery. The addition of the IR imagery confirms the identification and also provides status information on the victim. Fig. 2 depicts what appears to be a blank wall in the visible imagery. The wall is part of a box structure with a narrow opening that restricted our ability to enter without damage to the box. Current platforms would have likely passed this structure by without a second glance, particularly if the opening were not obvious or completely blocked. However, the IR imagery indicates that there is something in the scene. Overlaying the two images indicates that there may be a victim inside of the box.

In Fig. 3, we have overlaid icons representing the audio and the gas sensor data. The yellow horizontal bar and icon in the upper portion of the images represents the difference in amplitude between the left and right microphones. For example, the icon is shown more to the right of the display when there is more sound coming from the right. The blue vertical bar and icon along the right side of the images is a representation of the measured concentration of CO$_2$. A change in position indicates that the concentration went up or down as the platform was moved.

Processing of the audio and gas data revealed a few problems. The loudest sound heard by the microphones might not be the most important one to listen to. We will need to investigate source separation algorithms as well as consider adding more microphones. Further difficulty is encountered with the CO$_2$ sensors refresh rate. For the industrial version we chose, the refresh rate was on the order of seconds. This could result in missing CO$_2$ readings if the robot drove too quickly by a victim. However, even with these initial problems with the data, it is clear that the overlay display can show four distinct sensor modalities in a single image.

IV. DISCUSSION AND FUTURE WORK

Our results indicate that layered sensor modalities have potential to increase the likelihood of detecting hidden or non-obvious scene features. Current work has focused on demonstrating the utility of this layering technique. However, additional work is required to make this technique usable in a setting outside of the laboratory and to quantify the benefit of layering.

First, the layering will have to be performed in real-time. This layering will have to include registration of the various video images and any other sensors with associated directionality. This will have to be done on the platform to reduce the bandwidth of the transmitted data and will have to be accomplished at a nominal frame rate of 30 fps, although in some applications 15 fps is acceptable. We envision that software can do some of this but not at the 30 fps frame rate. Current television broadcast techniques use a combination of software and ASIC hardware to do image compositing in real-time and we believe that similar combinations of software and hardware have value in this application.

Second, the addition of layers must be selectable by the operator. The ability to turn on or off a given layer allows the operator to remove modalities that either do not contribute in a given situation or are confusing for certain types of operations. One example would be the display of CO$_2$ data. This data is useful when searching for survivors but does not provide much benefit when simply maneuvering. An added feature of this selectable approach is a modularity that allows the inclusion of additional modalities as new technologies become available or when extra information is required.

Third, the layered data should provide input into the autonomous control of the platform. Unless the platform is tethered, there is the likelihood that at some point during an operation the communications link between the operator and platform will be dropped. The layered data could help provide the platform with location awareness and environmental hazards sufficient to allow the platform to maneuver safely.
without operator control. Obviously, the visual display will not be of much assistance to the robot, but the sensor fusion during processing can be reformatted for the robot navigation algorithms.

Fourth, use of this type of data by experienced search and rescue personnel under controlled conditions is necessary.
Use of the technique in this environment will give us a quantification of the benefit provided by having a single display with multiple sensor modes displayed. This type of study will also allow us to understand further the limitations of the technique and provide opportunity to tune the displays.

Fifth, the problem of the choice of which sound is the loudest will have to be resolved. Heuristics can help but future work will also focus on using frequency analysis to help distinguish between victim emitted sound and environmental sound. Source separation methods could also assist in this process. Further work with the environmental sound can also be incorporated to assist in localization of the platform.

Lastly, the choice of gas sensors must be explored. We chose an industrial-use sensor with slow response rates but with a large dynamic range (0-2000ppm in 4-20mA measured current). Other sensors may have faster responses but with less dynamic range. Investigation into an optimal sensor choice will be necessary to move this capability into the field.

V. CONCLUSION

We have presented a technique for presenting multiple sensor modalities to an observer. The presentation is based on combining the sensor modalities into a single display in which non-visual modalities are displayed as iconographic visual cues. Preliminary results from the NIST Urban Search and Rescue Test Arena indicate that there is some utility in presenting the sensor data in this fashion.

ACKNOWLEDGMENTS

This work was supported by NIST 70NANB3H1116, NSF IIS-0308186, and a research seed grant from UMass Lowell. Thanks to Will Becker, Elena Messina, Monica Starnes, and Brian Weiss, all of NIST, for their assistance with data collection in the NIST Test Arena.

REFERENCES