Initial Design of a Multimodal Collaborative Mobile Application for Real Time Decision Making

Gregory Burnett[1], Thomas Wischgoll[2], Victor Finomore[1], and Candace Washington[1]
1. US Air Force Research Laboratory, 711th Human Performance Wing, Wright-Patterson AFB
2. Department of Computer Engineering and Computer Science, Wright State University

Abstract –*Mobile devices, smartphones and tablets, are continually expanding their computationalperformance capabilities through improved processing, interconnectivity network abilities, resource management and ease of use user interfaces. As such they are gaining interestas a means tosupport on-the-move remote collaboration for military personnel executing real time decision making tasks. This paper focuses on a software-based implementation of a prototype multimodal Android application that was designed to capture and disseminate real time battlefield perspectives to distributed entities. Moreover the mobile application enables remote experts to interactively collaborate through multimodal functionality to provide directives to the mobile userthat should be applied to the local scene. The design of themobile application interaction is discuss as well as the results from aninitialdemonstration where remote guidance was present to a mobile user attempting to defuse an improvised explosive device. Additionally, we report future implementation capabilities and projected military usage.*

Keywords – Multimodal, Mobile Computing, Remote Collaboration

1. Introduction

With changing environments and emerging unrecognizable threats, dismounted Battlefield Airmen (BA) need to carry advanced technologies to survive and effectively prosecute their missions. Requiring a high degree of situation awareness and the ability to multitask between various mission essential responsibilities, BA rely on mobile computing devices to stay informed of battlefield conditions. Within recent years, ground military research and development efforts have shifted from funding the minimization of rugged laptops and ultra mobile personal computers to leveraging mobiles devices such as smartphones and tablets [1]. These newer mobile devices continue to evolve at an unprecedented rate offering vast and ubiquitous capabilities. They allow users to perform tasks while on-the-move; as well as to be interoperable with heterogeneous distributed systems through a variety of communication channels. Mobile devices are increasingly becoming more user friendly, offering intuitive user interface controls and advanced features. Mobile device operating systems, such as Android, are improving power management and consumption of frequently used embedded features. Accordingly, mobile devices and their capabilities are being examined as potential decision making tools for military personnel.

A desired capability that dismounted warfighters seek to have is the ability to collaborate with non-collocated individuals on mission objectives and tasks. Through the use of mobile devices and their intrinsic communication and interface capabilities, real-time, on-the-move remote collaboration can be performed in dynamic battlespaces.

Remote collaboration on physical tasks is defined by Kraut et al. [4] to be: "*A general class of 'mentoring' collaborative physical tasks, in which one person directly manipulates objects with the guidance of one or more other people, who frequently have greater expertise about the task." (p. 16)* Task knowledge is important to enable remote collaboration; however to effectively collaborate, Clark et al [6] report there needs to be a mutual understanding between helper and worker to ensure *common ground*. This common ground can be achieved through various modalities. For example, when sharing a video feed of the active workspace, visual modal understanding can be achieved as the helper monitors the worker actions engaging in the task following an instruction. Alternatively, evidence of non-grounding could be observed if the worker hesitates to act out the remote instruction showing signs of confusion. Auditory, grounding can occur with verbal acknowledgments such as "okay", "got it", "unhuh". Conversely, audio modal non-ground can be heard through utterance such as "what", "huh", "I don't understand".

In this paper, we report the design and implementation of an Android mobile application that supports multimodal remote collaboration. We first highlight related work in

which we drawfeatures that will be incorporated in our application. These capabilities are implemented to achieve a higher common ground between non-collocated parties cooperatively working together on a task. Next, we discuss the methods and result from an initial demonstration of the mobile application, where participants worked jointly to defuse a simulated improvised explosive device (IED). Finally, future work to improve the prototype for military operations and potential military career fields that could benefit from real-time mobile collaboration in support of decision making is provided.

2. Related Work

Cooperative interactive systems between distributed parties working together to complete a physical task have been researched and implementedusing severalapproaches and apparatuses.

Kraut et al. [4]developed a wearable system consisting of head borne CCD camera, a VGA (640x480) display and microphone headset. The configuration had each worker don the system whichshared video and duplex audio between a helper and worker. The helper was able to monitor the video perspective of the worker's active workspace and use verbal instructions to guide maintenance procedure toward the task of repairing a bicycle. The authors' investigation concluded that field workers completed the task "more quickly and accurately when they have a remote expert helping them". Moreover, having a shared perspective positively influenced the verbal directives and feedback given between worker and helper.

Kirk, Fraser, &Rodden [3] designed a collaborative video/audio environment that sought out to address mixed reality ecology by conjoining two separate but similar workspaces into one hybrid workspace. The interactive system overlaid video capturedgestures and workspace elements of the helper onto the active workspace of the worker through the use of projectors. Creating a linked collaborative workspace, the helper could direct the workers actions through the use of simple hand gestures, audio commands, and/or both. Their results showed that task completion time and accuracy mistakes were less than voice alone when the helper used the combination of voice plus gestures.

Authors Ou, Fussell, Chen, Setlock and Yang's [2]DOVE (Drawing Over Video Environment) remote collaborative system facilitated a "remote helper to draw on a video feed of a workspace as he/she provides task instructions". The DOVE system supported both freeform annotation as well as gesture fitting recognition to generate a markup perspective shared to the worker. Results from their research suggest that markup capability "significantly reduces performance time compared to camera alone".

Interactive collaborative systems designed to support remote guidance evaluate their performance through various metrics. However, common metrics are task completion time, accuracy of task, response time, and mistakes made. Our paper focuses on these factors--assessing the usefulness of our prototype Android application as a tool to support real-time decision making. Moreover, cooperative systems utilize several modalities to achieve communication grounding between helper and worker.

Wickens and McClarley's [5] report that systems and interfaces utilizing multiple modalities are more advantageous to the user then those that do not. Their paper Suggests that multimodal interfaces allow their users to process different modality information concurrently for better cognitive understanding of the task at hand. For military use cases,cooperative interfaces leveraging various perception channels to communicate orders, instructions, battlefield information, etc.areessential. Warfighters often conduct operations in cognitively demanding environments that often require them to share their cognitive focus and attention across several events and stimulus happening concurrently.Additionally, auditory and visual distracters and/or masking are a battlefield constant. Special attention to designing independent or redundant multimodal capabilities is needed to provide information management to the warfighters.

3. Collaborative Prototype Design

Advancing user-centric cognitive interfaces for Battlefield Airmen (BA), researchers working in the US Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Interface Division began designing and implementing an Android mobile application for remote collaboration. Leveraging multimodal perception functionality, the mobile application sought to improve the warfighters capability by providing context rich information while supporting interactive collaboration of non-collocated parties. As depicted below, a cooperative application, running on a mobile device, could be utilized in the field to capture, disseminate, and interact with remote experts. These interactions can be support preplanned, dynamic, or time sensitive operations.

Figure 1: Mobile Device used to capture and share in field deployed perspective with remote experts

Drawing collaborative features from the highlighted related research approaches and other mobile application resources, our collaborative prototype identified the following set of capabilities for inclusion that military operators could benefit from:

1) Sharinglive video of active worker's workspace
2) Sharing full duplex audio between linked users
3) Support free form and predefined markup annotation on captured still from live perspective.
4) Adjustable transparency of overlaid markup image on-top of live workspace.
5) User configurable preview/live adjacent windows or merged preview and live perspective

3.1 Live Video Sharing

Streaming video of the active workspace has been shown to improve communication grounding and thus warranted its integration into our prototype application. Oursoftware design utilizestheintegrated camera of the mobile device platform or can connect to an off board camera through 802.11, Bluetooth, or USB. Once connection is made to the video capture device, an image buffer is used to store the cameras acquisition at anupper limit of 30 frames per second and adjustable down to a lower limit of 5 frames per second. The configurable sampling and pending transmission of the image is scalable to conserve power consumption. Prior to network transmission, the image buffer is compressed to an 800x600 jpeg to improve network utilization as well as maximize the receiving military parties' ability to process the image natively without preprocessing the data received.

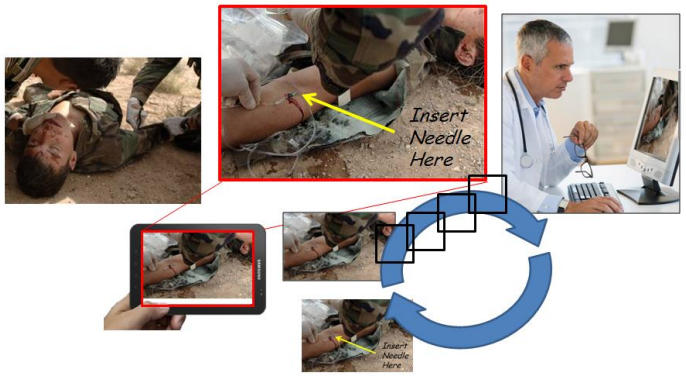

Figure 2: View Sharing between Helper and Worker

3.2 Audio

Audio communication was implemented using Voice over Internet Protocol (VoIP) functionality. The software supports live "hot" microphone as well as push to talk execution. Both audio input options were implemented to address constraints of military use casesthat could limit operators' available hands-on with the mobile device and for battery consumptionconsiderations. The application supports stereo or mono input and can be configured to use 8 or 16 bits per second at 11KHz, 22KHz, 44KHz, etc samples per second. The transmission and reception of network audio information are threaded to perform concurrently without delays.

3. 3 Markup Stills

Extending the visual modality collaboration capability, annotation of still images was implemented into our application. As Ou et al showed, cognitive performance and understanding improved through the use of markups.We designed functionality that allowed free-form and predefined elements to be captured and transmitted between helper and worker. The graphical inputs were collected

Figure 3: Markup Capability

through touch screen inputs on the mobile device display. When edits are completed the markup are merged with the still capture image for transmission. Image compression is applied and the resultant image is transmitted across the network.

3.4 Configurable Interface

Our mobile application supports two display configurations. 1) Full screen mode that renders the live perspective merged with a markup image on the same preview surface and 2) Adjacent mode that displays the live perspective and a markup image in separate preview surfaces.

Figure 4: Full Screen Mode and Adjacent Mode

The rationale behind having two display configurations can be drawn from the dependence on the cooperative markup received for a collaborative task. In the full screen mode, the markup can act as an exact guide or placement for the worker to perform task objectives. For example, the markup could identify a vein in a soldier's arm where an IV needle should be inserted. In the adjacent mode, the markup can serve more as a reference then a precise guide for a task. For example, place the tool on this shelf.

3.5 Transparency Adjustment to Markup

Markup images received from remote helpers will be displayed to the user and fused with the live perspective in the full screen mode. When attempting to perform a task, illustrated through markup context, the mobile user may choose to adjust the degree of transparency of the markup to better observe the active workspace. This capability was implemented in our application through intuitive touch screen actions. As depicted in figure 5, the user can adjust the transparency level applied to the markup simply by running a figure up or down the vertical dimension of the mobile device's screen. This dynamic ability to make the markup fade in and out of the live perspective serves to enable the user with an on demand overlay assisting collaboration.

Figure 5: Markup Transparency

4. Demonstration Experiment

To assess if cooperative interaction between distributed users can affectively be achieved on mobile devices, a relevant military task was evaluated. Improvised explosive devices plague military operations worldwide. With no constant design, IEDs have numerous wire configurations and trigger features. Defusing IEDs involve systemic sequential wire identification and disarming (cutting or rerouting wires) making the IED inert.

4.1 Participants

Twelve participants volunteered for this study, 8 men and 4 women, ranging in age from 23-30 (M =25) years. All participants had normal hearing and normal, or corrected-to-normal vision.

4.2 Design

A within-subject design was employed with four levels of Modality Interface (Audio, Video with Markup, Video with Audio & Video with Markup and Audio). The order for which each participant utilized a modality was controlled by counterbalancing the order so not to bias the experimental conditions. All participants took part in a training session to familiarize themselves with the task and devices. The participants trained defusing four IEDs per experimental condition. Participants were given the option for more practice trials; however, none of them felt the need for more. The four experimental conditions and IED configuration were randomized per participant.

4.3 Apparatus

Four simulated IEDs were used in the experiment. Each IED consisted of a clock, power source, control chip, and explosive charge containers as seen in Figure 6. There were nine wires on each IED, seven were active and two were distracters. The participants worked cooperatively with a remote confederate, who had detailed instructions for disarming each IED and experience communicating through the various multimodal communication capabilities. Participants used a Samsung Galaxy Tablet running our developmental Android application to interact with the remote confederate through a Wi-Fi connection. The Galaxy Tablet was mounted on a stand to allow the participant to freely use their hands, as seen in Figure 7. The remote confederate was situated in front of a workstation, which was isolated from the experimental area. The confederate's workstation allowed them to communicate via VoIP, capture, and annotate images from the participant's tablet to assist them in their task.

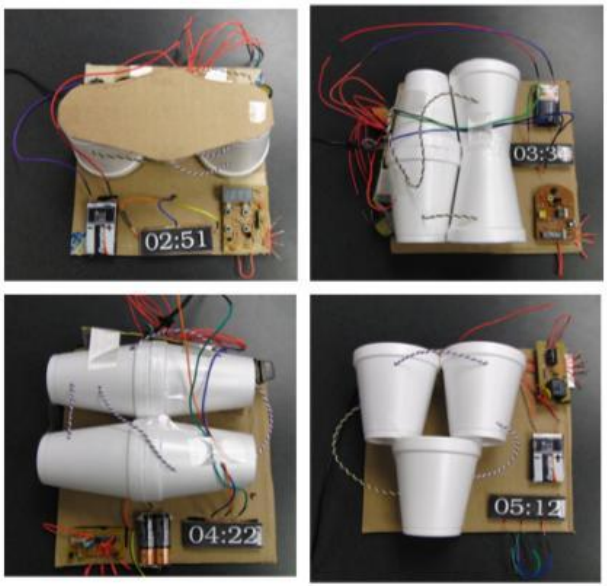

Figure 6: Simulated Improvised Explosive Devices

4.4 Procedure

Four conditions were evaluated. 1) Audio only where the confederate could not see the worker's workspace. 2) Video with Markup where the confederate monitored the workers workspace and provided markup directives. 3) Video with Audio where the confederate monitored the workers workspace and provided verbal directives. 4) Video with Markup and Audio where the confederate monitored the workers workspace and could provide directives through both markup and verbal interactions.

In the Audio condition, participants spoke to the confederate via VoIP where they had to describe the IED in order for the confederate to relay the sequence for which to disconnect active wires. The Video with Markup condition consisted of the confederate capturing a picture of the IED from the tablet's perspective then annotating the picture in real-time on their workstation. The annotated image, which showed the order of wires to disconnect, was sent to the participant to defuse the IED. The Video with Audio condition consisted of the confederate monitoring the participant's perspective while supplying verbal instructions to defuse the IED. The Video with Markup and Audio condition combined the Audio and Video conditions so that the confederate and participant were able to talk to each other as well as send annotated images.

For each condition subjects defused a unique IED. They were asked to complete the task as fast as possible without making any errors. A countdown clock was used to impose time pressure initially starting at one minute and decrementing each second.

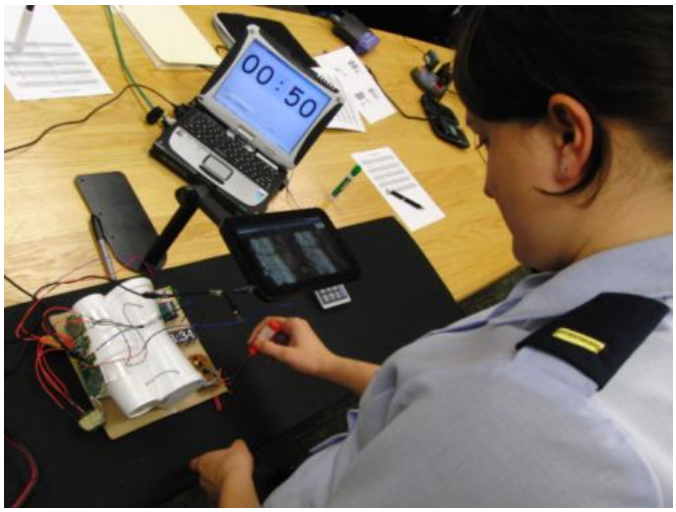

Figure 7: Participant diffusing IED with Tablet

5. Results

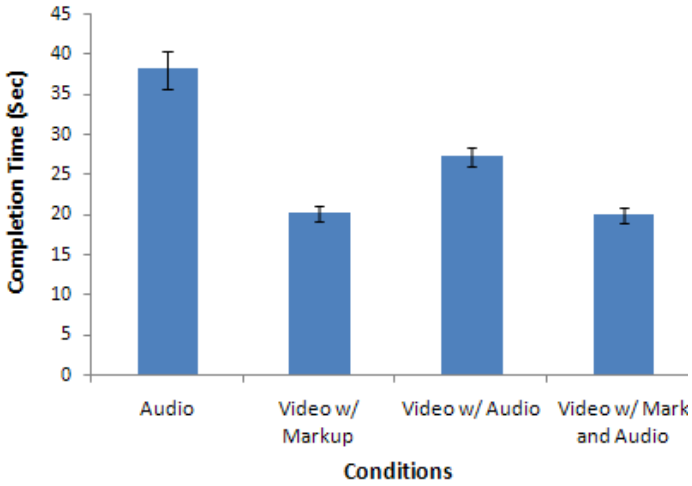

Table 8. Mean completion times for each of the four experimental conditions. Error bars are standard errors.

Mean task completion time and their respective standard errors for the four experimental conditions are displayed in Table 8. A 4 condition repeated measures Analysis of Variance of this data revealed a statistically significant main effect for conditions, $F(3,33) = 70.88$, $p < .05$. Subsequent post hoc Tukey-tests with alpha set at .05 revealed that participants using Video with Markup and Video with Markup and Audio completed the task statistically faster than the other two conditions but were not different from each other. The Tukey-test also found that participants using Video with Audio were faster than Audio alone.

6. Future Work

Mobile devices' form factor make them convenient for portability and on-the-move processing; however their compact size lends them to not remain static as users manipulate and interact with the device's context. When collaboration is performed through markup images the offset between the current live perspective and the captured markup perspective may slightly differ. In our current design when merging the live and markup in the full screen mode a ghost effect could be rendered if the two orientations do not align. To prevent this visually distracting effect, the ability to auto register the still markups with the active live perspective is desirable. Computer vision techniques can be incorporated into the cooperative Android application that performs feature extraction and image transformation on the annotated markups that orients them to the dynamic live image captured on the mobile device, as seen in Figure 9.

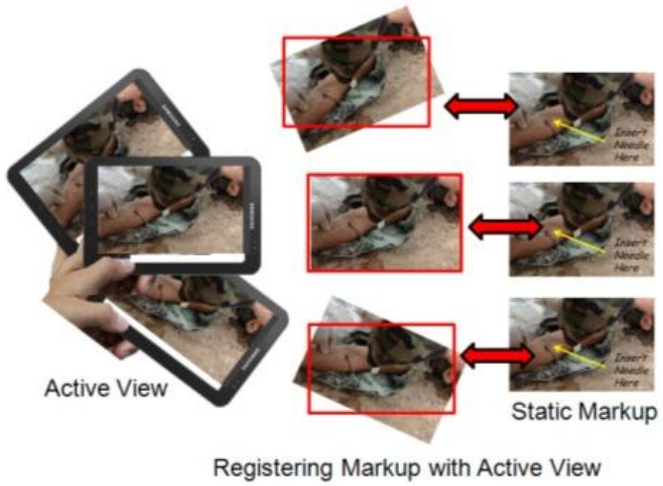

Figure 9: Image registration concept

Battery capacity is a major limitation of mobile devices. With regards to remote collaboration, power consumption and battery depletion prior to the completion of a cooperative task could be detrimental to the success of the task. Currently, it is left up to the mobile user to monitor their device's state of charge and adjust the mobile device's properties to prolong the runtime (e.g. reduce the screen brightness, disable updates, etc.). An additional improvement that we plan on incorporating into our Android application is a battery power status message. The message will transmit as additional information used in the collaboration between the helper and worker. With the mutual understanding of the current battery state of charge on the mobile device, both helper and work can monitor and negotiate features that could be reduced (frequency of active workspace transmission, sample rate of audio, on-board vs. off-board image registration, etc.) extending the runtime of the mobile device.

7. Conclusion

Previous work in computer supported cooperative work has shown that sharing multimodal information (video and audio) improves joint task completion. This paper reported the implementation of an Android cooperative collaboration application and its initial evaluation and demonstration use for defusing IEDs. Results from the experimentation showed that mobile device can support the communication capability to successfully complete a task jointly executed by a remote helper. The ability to share and annotate images was found to be the effective means to communicate directives. This initial investigation provided justification for further development of a mobile, multi-modal collaborative application for distributed operators.This application seeks to equip BA with a direct on demand link to SME to maximize mission effectiveness.

8. References

[1] Ackerman, S. (2011). Army Taps Android Phones for 'Wearable Computers'. *Danger Room* (6 Sept 2011), http://www.wired.com/dangerroom/2011/09/nett-warrior-smartphone/

[2] Ou, J., Fussell, S.R., Chen, X., Setlock, L.D., &Yang, J. (2003) Gestural Communication over Video Stream: Supporting Multimodal Interaction for Remote Collaborative Physical Tasks. In *Human-Computer Interaction Institute.*Paper 59.

[3] Kirk, D. S., Rodden, T. & Stanton Fraser, D. (2007) Turn It This Way: Grounding Collaborative Action with Remote Gestures. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, 28th April-3rd May, ACM: San Jose, CA, pp1039-1048

[4] Kraut, R. E., Fussell, S. R., & Siegel, J. Visual information as a conversational resource in collaborative physical tasks. *Human Computer Interaction,* 18, 1 (2003).13-49.

[5] Wickens, C. D., &McCarley, J. Applied Attention Theory, In Taylor & Francis, Boca-Raton, FL. (2008)

[6]Clark, H. & Wilkes-Gibbs, D. (1986) Referring as a collaborative process.*Cognition 22, 1-39.*