Multimodal Mobile Collaboration Prototype Used in a Find, Fix, and Tag Scenario

Gregory M. Burnett[1], Thomas Wischgoll[2], Victor Finomore[1], and Andres Calvo[3]

[1] Air Force Research Laboratory, 711 Human Performance Wing, WPAFB, OH
[2] Deptment of Computer Science and Engineering Wright State University, OH
[3] Ball Aersopace, Fairborn, OH
{gregory.burnett,victor.finomore,andres.calvo.ctr}@wpafb.af.mil,
thomas.wishgoll@wright.edu

Abstract. Given recent technological advancements in mobile devices, military research initiatives are investigating these devices as a means to support multimodal cooperative interactions. Military components are executing dynamic combat and humanitarian missions while dismounted and on the move. Paramount to their success is timely and effective information sharing and mission planning to enact more effective actions. In this paper, we describe a prototype multimodal collaborative Android application. The mobile application was designed to support real-time battlefield perspective, acquisition, and dissemination of information among distributed operators. The prototype application was demonstrated in a scenario where teammates utilize different features of the software to collaboratively identify and deploy a virtual tracker-type device on hostile entities. Results showed significant improvements in completion times when users visually shared their perspectives versus relying on verbal descriptors. Additionally, the use of shared video significantly reduced the required utterances to complete the task.

Keywords: Multimodal interfaces, mobile computing, remote collaboration.

1 Introduction

Battlefield Airmen (BA) serve a myriad of roles and responsibilities in an ever-changing austere battlespace. BA are equipped with wearable communication technologies to assist them in their mission objectives, enhance situation awareness, and provide interoperable means to communicate with distributed entities. Moreover, these mobile technologies are utilized to capture and disseminate battlefield information to distributed teammates to support dynamic decision-making (DDM).

Although BA are highly trained and possess a variety of skills, they can confront situations that are outside their expertise. Often the operator is left to complete the task on their own or must wait for an expert to arrive on scene, both unfavorable solutions in time-critical situations. The multimodal software approach described in this paper seeks to foster the ability of the dismounted operator to remotely collaborate with a subject matter expert (SME) on physical tasks. Kraut et al. [1]

D. Uhler, K. Mehta, and J.L. Wong (Eds.): MobiCase 2012, LNICST 110, pp. 115–128, 2013.

defines remote collaboration on a physical task as *"a general class of 'mentoring'… 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 [2] report there needs to be a mutual understanding between helper and worker to ensure *common ground.* This common ground can be achieved through the use of shared perspectives afforded by mobile device technology.

Advancements in mobile device hardware and software support DDM and foster remote collaboration. These improvements offer great potential to improve dismounted personnel's mission effectiveness, particularly during dynamic re-planning and mission execution. Mission performance and outcomes are often inversely correlated to time constraints resulting from real-time decision making [3]. Some examples include: survivability of time-critical causalities being attended to by infield medics through the guidance of a remote surgeon; battlefield repair of machinery by the end users being advised by a non-collocated mechanic; defusing improvised explosive device (IED) ordinance by untrained soldiers under the expert guidance of highly trained ordinance disposal personnel.

As military conditions and environments change so must the equipment used by BA. Within the last several years, there has been a research and technology development shift in wearable computing for combat ground personnel. Previous investments were spent miniaturizing rugged laptops and funding ultra-mobile personal computers development. Today, research focuses on leveraging and adapting emerging smartphone and handheld tablet devices for battlefield applications.

In addition to technical advancements, smartphones and handheld tablets are increasingly becoming more user friendly, with easy to use interface controls and advance features. Out-of-the-box, these devices' I/O mechanisms are designed to be interacted with while on the move. In addition, the packaging and flexible form factor of these devices enables them to be easily integrated into BA's personal battlefield ensemble. Another significant advantage is their lighter weight compared to previously fielded computing platforms. Lastly, mobile operating systems are aggressively improving power management with customizable power consumption features and settings. Leveraging the new mobile devices' vast and ubiquitous capabilities, research initiatives within the Air Force Research Laboratory (AFRL) are beginning to investigate their applicability to support multimodal mobile remote collaboration for BA and design intuitive interfaces to enhance mission effectiveness**.**

Smartphones and handheld tablets adaptation into warfighter ensembles depends on their intrinsic capabilities to function in multimodal communicative roles. It is often the case that dismounted BA must divide their cognitive resources across several tasks simultaneously. Audio, visual, and/or tactile perception channels may be masked or already occupied, rendering certain modalities ineffective for receiving additional information. Accordingly, special attention should be given in designing a mobile application that is flexible in its communication capabilities and modalities making it conducive to the extreme tempo and interaction of the dismounted military forces.

In this paper, we describe the implementation and design of an Android mobile application that facilitates multimodal remote collaboration. We first discuss previous related work and highlight features and capabilities that warrant inclusion into our application's implementation. These features seek to enhance greater communication grounding between teammates working cooperatively to complete a common task. Next, we report the methods and findings from a demonstration of the mobile application, where participants executed a joint find, fix, and tag scenario. Finally, future work and discussion sections are provided. These sections document enhancements to the developmental mobile application and discuss military benefits from real-time decision making capabilities supported by mobile collaboration.

2 Related Work

There are numerous technological approaches, multimodal apparatuses, and devices used to facilitate distributed collaboration. In the following section, we review several of these devices and highlight their design features and results.

Several researchers sought to use shared video perspectives to aid in collaboration of remote tasks. Kuzuoka et al [4] developed a spatial workspace collaboration system that enabled remote helpers to visually monitor the active workspace of the worker via a live video feed captured by a head-mounted camera worn by the worker. Moreover, the worker was able to visually receive guidance through a head mounted display that rendered the helper's physical actions recorded from a camera focused on the helper. Kraut et al. [1] leveraged a worker worn capturing setup consisting of a camera and a heads-up display, as well as audio headsets to provide verbal guidance to a worker. These remote collaborating apparatuses using shared visual information have been shown to decrease task completion times with greater task accuracy. Fussell et al [5] suggests that "visual information facilitates grounding or the development of mutual understanding between conversational participants."

Real-time markup annotation of shared images and live video between helper and worker has been shown to improve performance. Ou et al's [6] remote collaboration DOVE (Drawing Over Video Environment) system allows a remote helper to annotate a video feed with free-form and gesture-fitting markups while providing task instructions. Their findings reported markup capability "significantly reduces performance time compared to camera alone." Stevenson et al's [7] research utilized a combination of "on-video" and "in-workspace" annotation capability, where remote helpers could use illustrated guidance to direct the action of the worker. The use of annotation techniques reduced the spoken instructions into "spoken fragments like 'in', 'out', 'around', and 'here' as they drew" their remote directives.

Performance assessments of a collaborative system and/or application can be evaluated using several metrics such as instructional response time, task completion time, and task accuracy. Thus, the effectiveness of our Android application, designed to aid in the distributed decision making of cooperative teammates, will be assessed using these metrics.

The overarching goal of these research efforts and our Android application is to support remote collaboration through various mobile capabilities and modalities, establishing a shared mental awareness of the cooperative tasks in an effective and timely manner.

3 Design

The prototype Android application described herein is part of an on going program internal to AFRL's Human Effectiveness Directorate. The program develops advanced wearable information management and cognitive interface technology for BA. The mobile application's implementation was based on a user-centered design approach and implemented a multimodal, context-rich distributed system, which enhances interactive collaboration of distributed participants.

The primary components of the system are a mobile device and a personal computer. The user of the mobile device is called the worker, and the user of the computer is called the helper. Although the primary use case we envision consists of dismounted BA as workers and task relevant BA SMEs as helpers, the system can be used in other situations, such as among BA as shown in Figure 1.

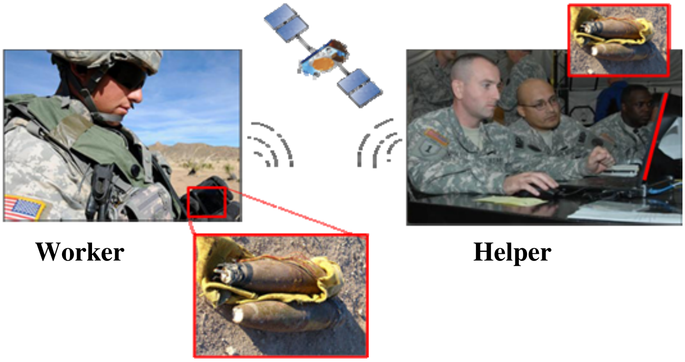

Fig. 1. Worker captures image of an explosive device with mobile devices and transmits image to Helper to obtain remote assistance

Drawing features from related work and understanding the dynamic mission conditions under which dismounted military personnel are required to execute, we formulated the following set of capabilities for our prototype mobile application:

- Sharing live video of the worker's workspace.
- Sharing full-duplex audio between linked users.
- Supporting markup annotation on captured still images.
- Overlaying annotated images on the worker's video feed with an adjustable transparency.
- Allowing the worker to switch between annotated image/live adjacent windows or merged annotated image and live perspective.

For the purposes of the prototype, we implemented data communication between an Android Samsung Galaxy Tablet and a personal computer using TCP/IP.

3.1 Sharing Live Video

Sharing a visual perspective of the active workspace is one of the fastest means to establish a common ground between cooperating individuals. Mobile devices are generally equipped with on-board cameras; however, depending on the circumstances, an off-board camera may be better suited for a collaborative task. Accordingly, our application was designed to accept a video capture device signal from either an embedded camera and/or an external camera tethered or wirelessly transmitting through TCP/IP. The degree of flexibility in video source(s) enables our mobile application to be scalable in order to address the demands and in-field capabilities.

Once a video capture device connection has been made live-video feed is transmitted to the helper as a series of 800x600 jpeg images at an average rate of 30 frames per second (fps). The mobile device displays a preview of the live-video feed as feedback for the worker. Implementing power conserving features, the frequency of image transmission is adjustable from 30 fps to 5 fps.

3.2 Sharing Audio

In addition to video communication, we implemented full-duplex audio communication across TCP/IP. The worker can choose to continuously transmit audio “hot mic” to the helper, or transmit audio only while depressing an external push-to-talk (PTT) button connected to the mobile device. Both input means were included in the design to address hands-free operations and for power consumption considerations. The mobile application can receive stereo or mono inputs and support a wide range of frequencies and sampling rates to accommodate the various military tactical communication headsets that may be used in conjunction with our system.

3.3 Still Image Capture and Annotation

Once the helper receives a live-video feed from the worker, he/she has the ability to capture a still image of the feed and annotate it with instructional content, as in Figure 2.

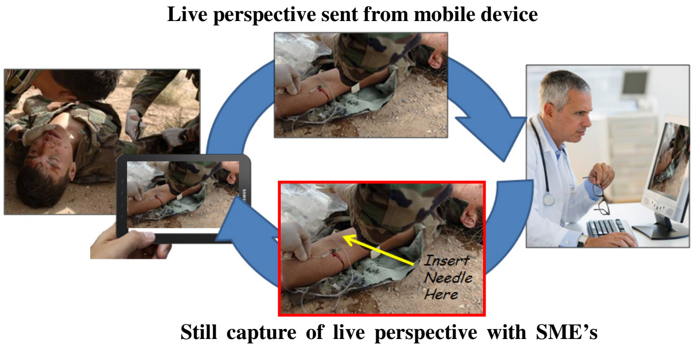
Live perspective sent from mobile device
Insert Needle Here
Still capture of live perspective with SME’s

Fig. 2. Helper captures image and annotates it

Additionally, the worker may want to focus the attention of the helper to a particular area in the live-video and can likewise initiate markup annotation on the mobile device. The mobile application supports both free-form markups and predefined objects and symbols assisting markup execution. Assigning one of the mobile device hot keys, configurable in the source, to capture an image from the live-feed, the worker can then use touch screen interactions to create his/her desired additions to the image prior to transmitting it to the helper (see Figure 3).

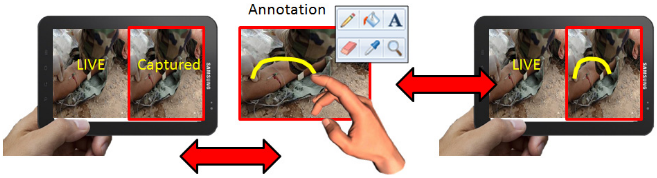

Fig. 3. Worker annotates image

3.4 Adjustable Display Modes and Markup Transparency

Annotated images are presented to the worker in either of two display modes: full-screen or split-screen, as shown in Figure 4.

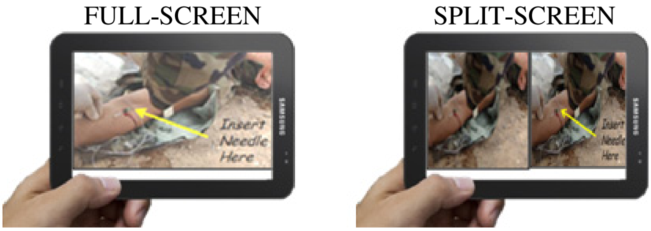

Fig. 4. Mobile application display modes

In split-screen mode, the worker sees the live-video feedback on the left half of the screen and the annotated image on the right half of the screen. In full-screen mode, the worker sees the live-video feed with a translucent overlay of the annotated image on the entire screen. The worker is able to toggle between split-screen and full-screen modes by pressing a button in the mobile device's options menu.

In full-screen mode, the worker adjusts the transparency of the overlay by using a pan gesture on the left half of the screen, as shown in Figure 5. Panning up reduces the transparency, and panning down increases the transparency. With these gestures, the worker can quickly set the transparency of the overlay to a level suitable for the current task. Note that the actions we chose to control the overlay's transparency level and toggle between display modes do not occupy any space on the user interface (UI) that would be used by the live-video feed or the annotated image. Thus, these controls do not get in the way of the live-video and annotated image.

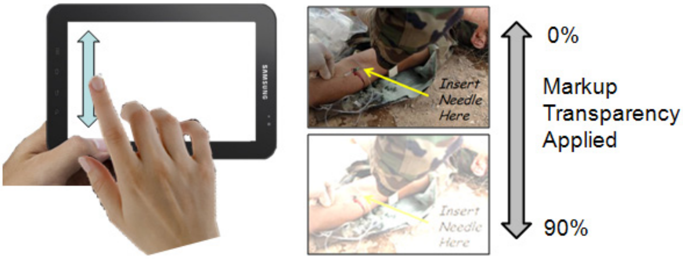

Fig. 5. Markup transparency setting control

Workers can observe the annotated image in a way that they find suitable for the current task by switching between display modes and adjusting the overlay's transparency. For example, consider a scenario where the worker is a BA and needs to place an intravenous (IV) needle in a soldier's arm. The worker shows the helper, who has medical training, the live-video feed of the soldier's arm. The helper captures a still-image and annotates the best place to insert the IV needle. The worker can then set the display mode to split-screen to see where he/she needs to place the IV. Alternatively, the worker can set the display mode to full-screen with a medium transparency and aim the camera at the soldier's arm to obtain a precise guide for the task.

3.5 Data Exchange

Interfacing with tactical heterogeneous systems, the multimodal data captured by our application was not preprocessed with video/audio compression prior to transmission. Network and data optimization are handled by an intermediate network node between the cooperative helper/worker pair communicating with each other.

4 Evaluation

An initial demonstration involving a find, fix, and tag task was conducted to assess the extent to which the developed Android application supported remote collaboration. Participants communicated with a remote expert using various modalities to complete the evaluation task. Task components involved: 1) identify a specific individual from a crowd of people, 2) align aiming device on identified individual, and 3) initiate a tagging sequence. The modality interfaces investigated were Audio, Video with Markup, Video with Audio, and Video with Markup and Audio.

4.1 Participants

Eight military and four civilian participants volunteered for this study (eight men and four women) ranging in age from 23-30 (*M* =25) years. All participants had normal hearing and normal, or corrected-to-normal vision. Additionally, all participants had prior training and experience handling a rifle. The participants collaborated with a remote confederate, who knew the sequence and identity of the individuals being tagged.

4.2 Experiment Design

A within-subject design that was balanced using a Latin-square procedure was employed with four levels of Modality Interface (Audio, Video with Markup, Video with Audio, and Video with Markup and Audio). All participants took part in a training session to familiarize themselves with the task and devices. The participants trained by communicating with the remote confederate and marking targets of interest with an AirSoft M-4 rifle 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 virtual target configurations were randomized per participant.

4.3 Apparatus

Each participant (Worker) used an affixed pivoting AirSoft M-4 Rifle with a camera attached to the forward barrel as shown in Figure 6.

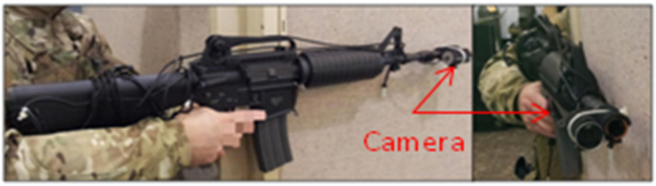

Fig. 6. Rifle with attached camera

Participants were instructed to stay behind a partition wall, which blocked their line of sight to the active scene, and utilize the rifle mounted camera's perspective for the task, as seen in Figure 7. The partition wall was positioned in front of an 8'x10' projection screen that rendered a virtual scene consisting of a gathering of 12 potential targets of interest.

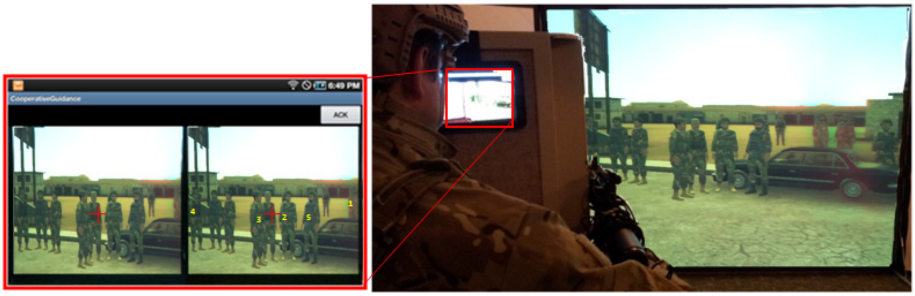

Fig. 7. Structure and experiment scene

The rifle/camera provided a live video feed to a Samsung Galaxy Tablet running our developmental Android application. The Tablet was stationary mounted to the partition wall allowing the participant to freely use their hands, as seen in Figure 7.

The remote SME (Helper) communicated with the participant through the tablet running the collaborative Android application through a Wi-Fi connection. They were situated in front of a workstation, which was isolated from the experimental area, as shown in Figure 8. The SME workstation allowed the cooperative pairs to communicate via streaming audio as well as capture and annotate still images from the participant's tablet. The SME used this tool to direct the participant in finding and tagging the hostiles in a specific order.

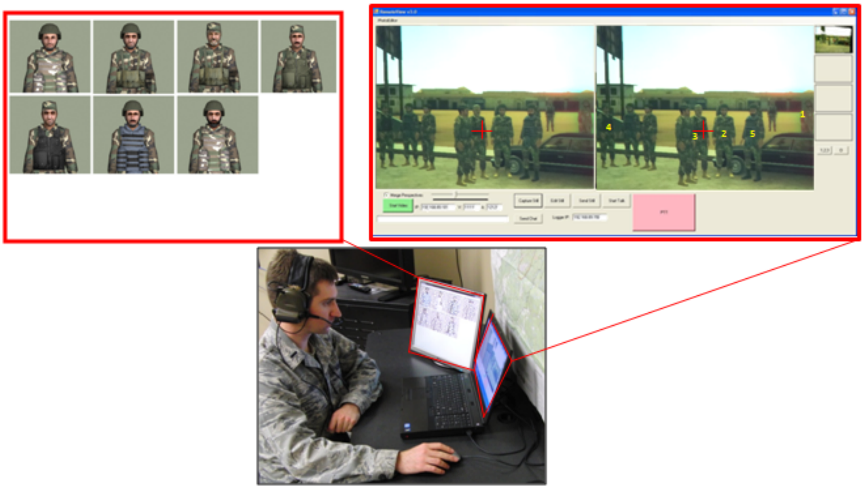

Fig. 8. SME collaborative workstation

4.4 Procedure

The four conditions that were evaluated included:

1) Audio only, where the SME could not see the participant's camera perspective and thus could to give directions via voice commands only.
2) Video with Markup, where the SME could monitor the participants' view from the camera mounted on the rifle and was able to annotate the still images, providing directives on the tablet.
3) Video with Audio, where the SME was able to monitor the participants' perspective and provide verbal directives.
4) Video with Markup and Audio, where the SME was able to monitor the participants' perspective and could provide directives through both markup and verbal interactions.

In the Audio condition, the SME had to verbally describe to the worker the characteristics of the individual that required tagging. The SME description of the individual started with a clothing description, an indication of facial hair, and whether the individual was wearing anything on their head. The Video with Markup condition consisted of the SME capturing a picture of the participant's perspective from the rifle mounted camera then annotating the picture in real-time on their workstation. The annotated image, which showed the order of individuals to be tagged, was sent to the participant to initiate the tagging action. The Video with Audio condition consisted of the SME monitoring the participant's perspective while supplying verbal instructions of the individual to be tagged. The Video with Markup and Audio condition combined the Audio and Video conditions so that the SME and participant were able to talk to each other as well as send annotated images.

For each condition, participants tagged unique individuals. They were asked to complete the task as fast as possible without making any errors.

5 Results

Mean task completion time and standard errors for the four experimental conditions are displayed in Figure 9.

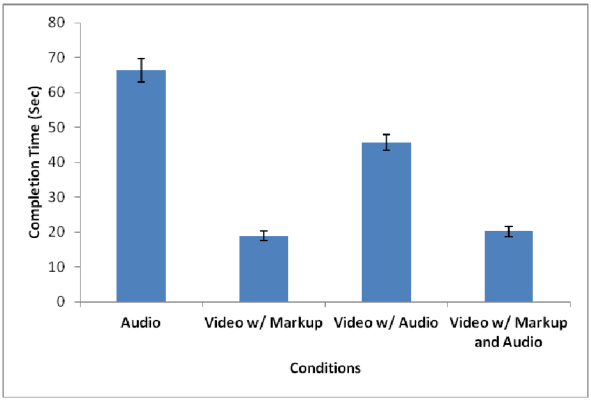

Fig. 9. Mean completion times for each of the four experimental conditions. Error bars are standard errors.

A four condition repeated measures Analysis of Variance (ANOVA) of these data revealed a statistically significant main effect for conditions, *F* (3,33) = 70.41, *p* < .05. A subsequent post hoc Tukey-test 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 conditions but were not different from each other. The Tukey-test also found that participants using Video and Audio were statistically faster than Audio alone.

Mean accuracy and standard errors for the four experimental conditions are display in Figure 10.

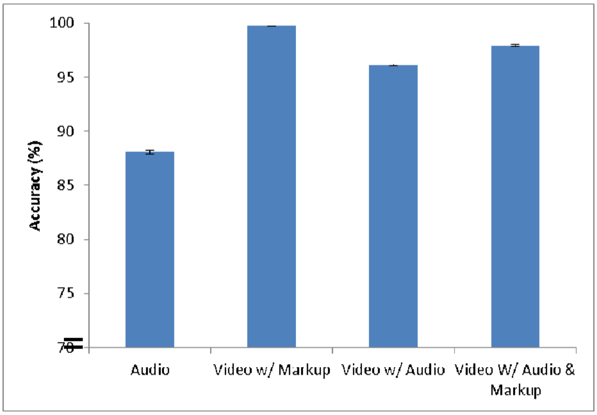

Fig. 10. Mean accuracy for each of the four experimental conditions. Error bars are standard errors.

A four condition repeated measures ANOVA was performed on these data and revealed that the mean accuracy values in the four conditions did not statistically differ from each other, *F* (3,33) = 2.24, *p* > .05.

Additionally, we examined the degree to which the interfaces provided affected the total verbal communication time. It was found that the style and amount of verbal information relayed between cooperative pairs differed when a shared visual perspective was available. Figure 11 shows the mean voice usage times the remote SME required to achieve common ground positively identifying the experimental targets. A t-test revealed that the Audio condition required more communication time then the Video w/ Audio, t (7) = 4.27, p < .05.

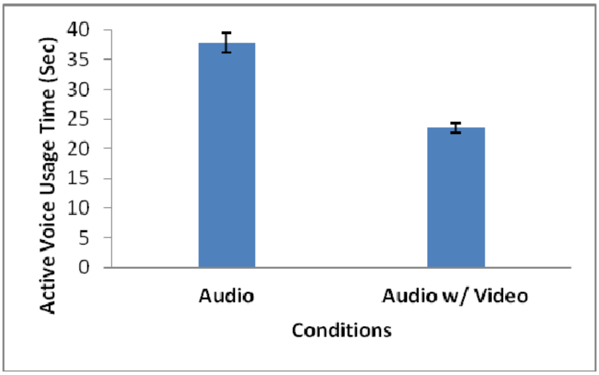

Fig. 11. Comparison of active voice usage time of audio conditions

6 Discussion

Our results are in line with other research evaluating the utility of various multimodal displays in remote collaboration tasks. Our data suggest that multimodal remote collaboration can be effective on mobile devices, such as smartphones and handheld tablets. Our findings highlight the fact that, when using shared visual perspective, cooperative teams can complete distributed physical tasks quicker and with the same level of accuracy as the other experimental modality conditions.

Not only did having a shared visual perspective result in a faster convergence of understanding, but it also had an impact on the style of the verbal directives. For example, in the Audio only condition, with no shared visuals, the remote SME's verbal directives were descriptive describing the appearance of the individual of interest (i.e., "the first guy has no hat [pause] white beard [pause] and a gray shirt", "the next guy has a brown hat [pause] small black beard [pause] and a white shirt"). Alternatively, in the Video with Audio condition, the verbal directives provided contextual information on the targets' location in the shared visual scene (i.e., "all the way to the back next to the car [pause] that one [pause] yep", "the fifth one to the right"). Moreover, in the Video with Audio condition, the remote SME leveraged pronouns such as "that one", "him", "next one" to convey and direct the worker's aim towards the correct target. The descriptive and contextual information we experienced is similar to the classification of utterance ideas of *Referents* and *Position* presented in Kraut et al [8].

Results also revealed that participants were quicker at completing the task when the helper sent annotated images than when they gave only verbal directions. This could result from the fact that both the helper and participants simply had to look at the image to interpret the directives rather than interpreting the spoken message. This finding supports the famous adage "A picture is worth a thousand words".

7 Future Work

Network/Visual optimization can further enhance our prototype Android application to be executed in bandwidth-limited environments. As we highlighted in the discussion section, collaborating participants preferred to leverage the visual modality to relay directives when given the option. Our current implementation redraws a new 800x600 jpeg still image to the mobile device's display for every helper's markup content. For greater efficiency, we may choose to implement region of interest redraws/image invalidations depicted in Figure 12. This can improve the system in

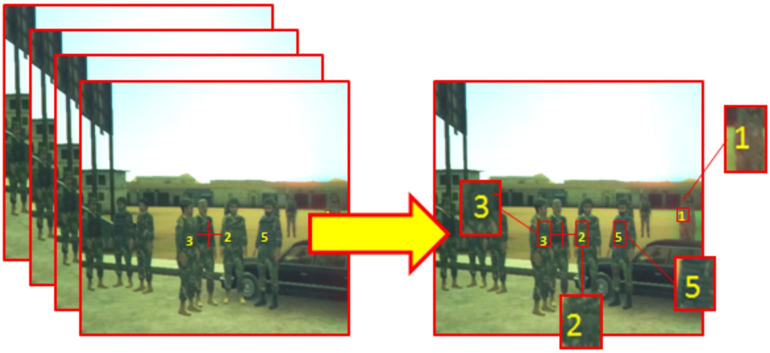

Fig. 12. Region of interest refresh

two ways: 1) It will improve bandwidth efficiency by only transmitting the bounding contexts contained in the markup and 2) It could be used to draw attention to the helper's markup with a distinguishing highlighting color indicating to the worker a change in information.

As the typical use case for military operations is not in a static position or environment, an evaluation of our developmental Android application's performance is warranted when worn in a new location and orientation rather than mounted to a secure object. Recent developments in low-profile wrist mounts (see Figure 13) offer potential test wearable configurations that may be conducive to dismounted on-the-move operations.

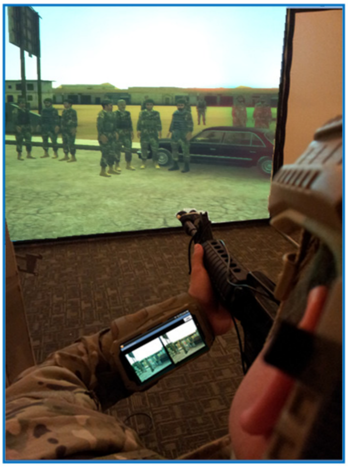

Fig. 13. Wrist mounted mobile devices

8 Conclusion

Computer supported cooperative work and team collaboration research has shown that sharing multimodal information (video and audio) enhances cooperative task completion. This paper reported the implementation of an interactive Android collaborative application and an initial evaluation executing a find, fix, and tag scenario. Results from the experiment showed that mobile devices can support the communication capability to successfully complete a task jointly performed by a worker and a remote helper. The cooperative use of annotated images proved to be the most effective means to relay instructional directives. This evaluation provides justification for continual development of mobile, multimodal collaborative applications for distributed operators. Such applications can enhance mission effectiveness by providing BA with a means to more effectively collaborate and coordinate critical military tasks with a remotely located expert.

References

1. Kraut, R.E., Fussell, S.R., Seigel, J.: Visual Information as a Conversational Resource in Collaborative Physical Tasks. Human Computer Interaction 18, 13–49
2. Clark, H., Wilkes-Gibbs, D.: Referring as a Collaborative Process. Cognition 22, 1–39 (1986)
3. Brehmer, B.: Dynamic Decision Making: Human Control of Complex Systems. Acta Psychologies 81(3), 211–241 (1992)
4. Kuzuoka, H., Kosuge, T., Tanaka, K.: GestureCam: A Video Communication System for Sympathetic Remote Collaboration. In: CSCW 1994, pp. 35–45. ACM Press, New York (1994)
5. Fussel, S.R., Kraut, R.E., Siegel, J.: Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In: CSCW 2000, Philadelphia, PA (2000)
6. Ou, J., Fussell, S.R., Chen, X., Setlock, L.D., Yang, J.: Gestural Communication over Video Stream: Supporting Multimodal Interaction for Remote Colloborative Physical Tasks. In: International Conference on Multimodal Interfaces. ACM Press, Vancouver (2003)
7. Stevenson, D., Li, J., Smith, J., Hutchins, M.: A Collaborative Guidance Case Study. In: AUIC 2008, Wollongong, NSW, Australia (2008)
8. Kraut, R.E., Darren, G., Fussell, S.R.: The Use of Visual Information in Shared Visual Co-Presence. In: CSCW 2002, New Orleans (2002)