Collaborative Mobile Video Conferencing Supporting AR Stabilized 3D Hand Gestures

Kevin Ta
University of Calgary
kta@ucalgary.ca

ABSTRACT
Modern video conferencing tools can be used for remote collaboration, but causes confusion when gestures are used in remote-expert scenarios. Experts often use hand gestures as shortcuts to describe what they mean; however, because the cameras do not capture these gestures in their intended context, communication is then focused on disambiguation rather than the task. Prior work has often captured either the workspace or gesture in 2D which is only particularly suited for tasks on a flat plane. However, these systems cannot capture or present gestures in a meaningful way when objects become occluded by depth. Real world tasks are often three dimensional which requires users to take on different perspectives and make gestures with depth. I built a prototype system to investigate the use of making and perceiving hand gestures in stereoscopic 3D for remote-expert scenarios. In this approach users can make gestures that can communicate position and scale within a 3D space. In addition, users can work at any angle to the workspace, moving around as necessary to overcome occlusion. A pilot study was conducted to determine design factors that affect user behavior through 3D stereoscopic capture and display of hand gestures in similar real-world tasks. My findings suggest that future work should focus on improving the prototype's fidelity and examining a wider variety of gestures under different rendering scenarios.

Author Keywords
Hand gestures; video conferencing; remote assistance; augmented reality; stereoscopic 3D; head mounted display.

ACM Classification Keywords
H.5.3. Information Interfaces and presentation: Group and Organization Interfaces – CSCW

INTRODUCTION
Today's video conferencing tools (e.g., Skype, Hangouts) are ineffective in communicating gestures for use in collaborative tasks. Consider the situation in which a remote collaborator is helping a friend diagnose and repair a computer. When providing assistance, it is difficult to refer to the many components within the computer. If instead the users were collocated, they would simply use a pointing gesture. However because these gestures are lost given the capture tool (i.e. webcam), they continue to point with their hands as though their partner could understand what they meant [4].

Unable to communicate clearly, workers waste time on trying to disambiguate what their partner is saying rather than focusing on the task. This is better dealt with hand gestures that can communicate rich and subtle information succinctly [6]. For example, consider an artist showing their apprentice how to use different hand drawing techniques. The way the artist grasps, angles, and stroke the paper with the pencil are all easily communicated with a single hand gesture concurrently. Furthermore the artist can verbally annotate this gesture to further clarify their technique. Contrast this to giving only verbal instructions which must communicate all three pieces of information sequentially to achieve the same effect.

However current gesturing systems are insufficient for real world tasks as they are not suited for 3D tasks. Prior work have focused on collaboration in fixed perspective scenarios [e.g. 5,7,8,9,10]. Users are often placed in shared workspaces and are forced to view and interact on workspaces whose screens lack stereoscopic displays. This flattens the workspace into a 2D plane which is unsuitable for working with 3D objects that require depth to fully analyze the object. Such objects need to be viewed at various angles in order to overcome occlusion.

Even when users are able to view the workspace at different angles, it is difficult to communicate using only 2D representations of gestures. Using only 2D gestures is difficult to perceive depth since it is visually cued by scale. In the case of a computer repair, some areas are difficult to reach because they are often occluded by components (e.g. cabling). One utility of hand gestures is to describe how to approach an object by describing a route to reach that object. However this is difficult to do with only 2D representation because depth is visually cued by scale which is difficult to distinguish if the change is subtle. Hence it is not enough to capture 3D gestures, but there is also the need to display them in a stereoscopic display so that depth is not cued by scaling the gestural representation.

To address these problems I built a video conferencing system that integrates 3D gestural communication. I built a prototype that captures 3D hand posture and renders them in stereoscopic 3D in another person's display. The recipient (worker) of the render wears a head mounted display (HMD) to allow both participants to use both their hands. Additionally, the worker will be able to see the hands with depth information which allows more complex maneuvering and perception of hands.

By capturing and displaying 3D hand gestures my prototype attempts to broaden the set of tasks to physical 3D objects.

Where prior systems have focused on a fixed shared workspace, my system allows users to work at any perspective and position. In this manner, users can look around objects and overcome occlusion due to depth. Additionally, the scale at which the gestures are made can be used to communicate and gauge relative distances between objects in 3D space.

The prototype attempts to explore hand render positioning using AR. Realistically a collocated collaborators' hands are not always placed directly in the front, but should be placed where they are the most meaningful to the worker (e.g. across a table). Using AR will allow the worker to place a physical object, look at it with a mobile camera, and see the 3D hands as though it were placed in the real world.

To measure the effectiveness of this tool, I ran a within subjects pilot study with pairs of participants to evaluate the interface of this system and determine design factors for such a system. The pairs performed a series of instructions that involved orienting 3D pieces; analyzing the workspace at various angles; and making precise hand postures to communicate instructions. Three gesture display configurations were conducted to evaluate the task design's ability to reveal differences between the systems.

I found that all participants made rapid and highly complex hand gestures that my system simply could not capture. In addition, there were a number of technological issues that hindered the prototype's effectiveness. Hence my work contributes several design factors for future work developing mobile video conferencing tools supporting collaborative hand gestures. This includes improving hand capture fidelity; the size and location of the interaction volume; and task design to examine wider array of gestures under different rendering scenarios.

RELATED WORK

Prior work has greatly influenced the design of my prototype. I will now discuss key related works that have been addressed in the design of my prototype.

Utility of Hand Gestures

Researchers have found hand gestures to be an important means of communication when collaborating. In a block construction task, Jones et al. [4] found that when using video conferencing tools such as Skype and Hangouts to collaborate, participants often used their hands to point at the screen to refer to pieces, but despite their pointing these gestures are not captured with their hands on the referred object. This is a cause for confusion as collaborators try to disambiguate what their partner is trying to refer to.

In addition, participants would often use deictic references such as "this" and "over here" relied on their hands as pointers (e.g., [4,11]). These references do not make sense without their partner pointing to the object at the same time. Hand gestures are rich communication tools, however when their intended messages are not captured we need to spend time to reestablish common awareness between both users [7].

There are instances where audio and/or annotations only communication cannot efficiently transmit information. The *reference space*, according to Buxton [1], is a space in which a helper can use body language to reference things in the workspace. In this space "one can sense proximity, approach, departure, and anticipate intent" which are all information channels that hand gestures can communicate. However, confusion arises when hand gestures to refer to real world objects that are not effectively framed or supported well.

Role of Depth Perception

Current related literature have focused on developing collaborative systems that force users into a constrained and shared workspace. In these situations a user is often placed in front of a monitor or display table [e.g. 5,7,8,9,10], and are forced to view interactions with the system along a flat surface. It is either that the system captures the world perpendicularly (thus removing depth), or displays without a stereoscopic display. Both cases lack depth which simplifies the space of tasks to being on large flat surfaces and manipulating 2D objects. It is easy to discount tasks in which users are faced with a 3D objects where details may be occluded by depth.

In addition, with the emergence of mobile video conferencing systems, capturing the workspace will often not be at ideal camera angles. Hence we need to move away from fixed and shared workspaces to explore collaborative systems that utilize depth for both capture of gesture and perception of workspace.

Stabilized vs non-stabilized objects

Annotations themselves are often much easier to work with if they are attached to objects in the real world [2,6]. In a computer repair scenario, users may need to fetch various items outside the workspace. As users return to the workspace, the annotations lose their original contexts. It can be problematic when users may need to refer back to their annotations as they have to realign their annotations to their original contexts in order to make sense of them. Prior work has looked into stabilizing annotations by attaching them to real world objects, which users have found to be useful [2]. Likewise by introducing world stabilization to hands, users will be able to make gestures without needing to maintain alignment. This is important for hand gestures because it may ease users as they try to precisely mimic their partner's hand movements while maintaining a still scene to make gestures in.

Supporting Representations of Multiple hands

While prior work has worked with multiple hand capture systems, many have yet to consider gestural presentation in mobile video scenarios. The BeThere system [11] captures only a single hand and requires both users to operate a camera. This makes it difficult to accomplish tasks that may require the use of two hands (e.g., playing the drums). And

even in cases where both hands are captured (e.g., [3,7,11]), hands and arms may collide the workspace [8] due to the lack of height information [5]. However the utility of dual handed gesturing is valuable as supporting them will allow users to make an extended variety of gestures such suggesting a method for combining two objects together [7]. Then the challenge is representing multiple hands in the in mobile video conferencing scenarios. These scenarios introduce height, perspective, and camera mobility as design challenges to address when presenting hand gestures to a user remotely. Hence a new approach is needed to represent hands in a way that is distinguishable for a remote user and supports multi-handed gesturing.

PROTOTYPE

System Design

When designing a system to support gesturing in remote mobile collaboration, gestures need to be presented in 3D and be able to support multiple camera angles. Unlike prior work that has worked in fixed camera scenarios, having a mobile camera affords the ability to look around 3D objects. However gestures will need to be supported in a way that is not occluded or distorted while looking around an object. To solve this, my system captures hand posture in 3D and displays them as a 3D model. Capturing 3D hands gives gestures the ability to convey depth information. Using depth allows hands to maneuver around objects and not be bound by a fixed 2D plane. It is also clearer to see depth as it is not cued by scaling the object, but by using 3D perspective rendering.

However to address the issues of maintaining context for the hand gestures, I have integrated an augmented reality stabilization for hand gestures. Prior work have observed loss of context effects from annotations and have developed AR stabilization methods [2] to maintain context for such objects. Gestures will need to maintain their contexts since they are less persistent than annotations. Using augmented reality stabilization, gestures are attached to a fixed point in the real world using an AR tag. Users can then expect the hand to remain in one place while their partner is operating the camera. This is important because gestures need to maintain their gestural context that will otherwise be lost when a mobile camera moves. Additionally this helps users mimic their partner's hand gestures while being able to move the camera freely.

To address the difficulties of operating a mobile camera, a head mounted display was used to display the gestures onto a working user. This setup allows users to utilize both their hands while allowing the freedom to look around comfortably. The advantage is being able to make multi-handed gestures that were previously limited to fixed camera scenarios. Combined with AR stabilization, users can comfortably look around without losing the contexts of their gestures.

In addition, using a head mounted display enables the use of a stereoscopic display. A stereoscopic display offers the ability to perceive gestures with depth more readily than using perspective 3D. Using depth as a cue, users will be able use gestures as a way to judge sizes and scales that are otherwise impossible using 3D perspective rendering on a 2D display.

Adding a shadow helps create a point of reference to allow users to sense height. At arbitrary camera angles, a 3D hand on a screen appears to float in midair. By adding a shadow to the hands, the system is able to cue users about the relative height off a table surface. A point of reference is important because it grounds the user's sense of height as the user looks around an object. This makes it clear to where the hands are floating and how far the hand is away from the user. This is done even in the stereoscopic case since the prototype is limited to a 2D video background due to the single camera.

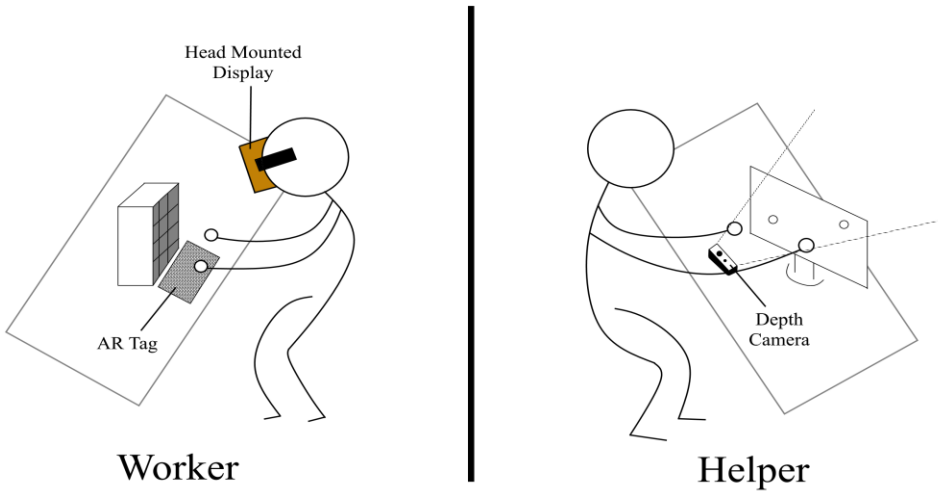

Figure 1. Prototype Schematic. A Worker (left) receives hand gestures via a head mounted display and looks at the structure using the AR Tag. The helper (right) makes hand gestures in front of a monitor using a depth camera to assist the worker.

The capture system was designed to create as much overlap between the user's hands and virtual hands. Prior work have often situated hand capturing systems over top the workspace so that the hands being captured are aligned with the remote representation. When collaborating remotely with a partner, users generally made pointing gestures front a monitor. The prototype is designed to take advantage of this experience and was designed to have the hand capture volume directly in front of the monitor. This aligns with being able to make gestures in front of the screen so that they projected directly to their partner's view.

Implementation

A prototype was built using the Unity3D game engine and the Vuforia Augmented Reality toolkit. The system captures hand posture from a helping user through an Intel RealSense depth camera. A working user then receives these hand gestures in a stereoscopic head mounted display. Using augmented reality tags, the system automatically renders the hand to a fixed position within the real world.

Hand Capture

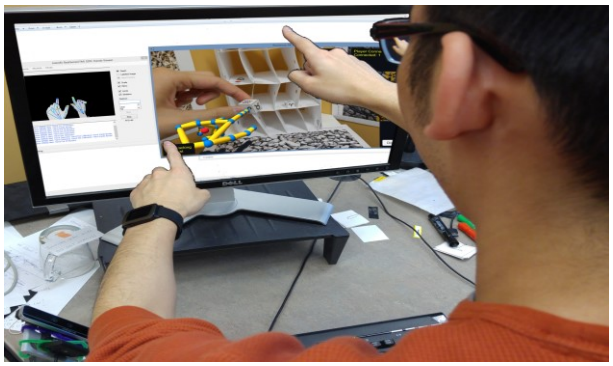

Figure 2. A helper making gestures to assist their partner. A diagnostics screen is presented (left). A hand capture camera (bottom left) is mounted at a 45 degree angle.

The system is able to track two hands simultaneously including their fingers and render them remotely to another device. This is achieved by mounting an Intel RealSense hand capture camera at 45 degrees facing upwards towards the ceiling capturing the palm of a user's hands. Using this method, the camera is able to sense more open palm orientations and pointing gestures than by mounting the camera facing perpendicularly to the user's palms. The camera is then set a fixed position away from the monitor so that hands are only captured while in the space in front of the monitor.

Hands are then displayed on a head mounted display on Google cardboard. The system has been calibrated enough to accurately render the hands so that remote users will see hands precisely overtop the intended object. This is consistent with Kirk et al. [7] to promote as much overlap of awareness as possible. Even when the hands are outside the camera view, the awareness that the hands cannot be seen is shared between both users.

The hand capture system captures hands in a special space in front of the monitor which I define as an interaction volume similar to what Sodhi et al. [11] describes. An interaction volume is the space in which the hand capture system works ideally in. This can be arbitrarily defined which I have calibrated it to the space in front of the helper's monitor. This is consistent with the literature as participants tend to make gestures directly in front of the monitor [4] aligning with past experiences.

In conjunction, a camera status screen is also presented to the helping user to diagnose weather their hands are being recognized by the system or not. This aids the helper as the camera may not recognize the hands if they are too close, too far, or in an unrecognizable posture.

Augmented Reality Stabilization

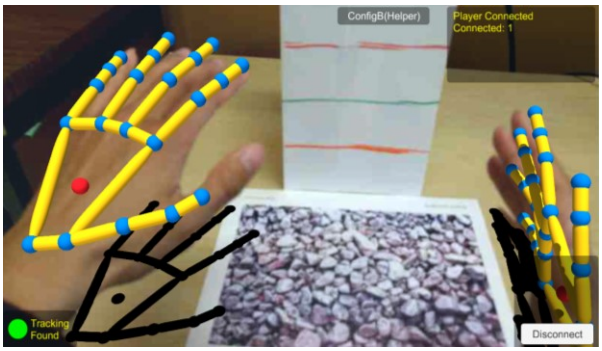

Figure 3. A worker mimics gestures communicated by a remote partner's hands through a 3D skeleton model. The augmented reality tag is the stone texture.

With the hands accurately captured and rendered, an Augmented Reality stabilization has been implemented to stabilize the hand render to the real world. Using the Vuforia toolkit, an image target was defined as a cylinder marked with a unique texture (see Figure 2.). The system can detect this texture to determine the cylinder's orientation and position which can be changed in real time. The hands are then repositioned to the target which makes them appear as though they are attached to the target. Rotating the cylinder target changes the direction in which the fingers are facing. These actions are reflected in real time to both collaborators using the system.

However for the hands to appear stabilized to the real world the AR image tag must be visible in the camera view at all times during the collaboration. Additionally, the worker wears a head mounted display, they can look at the workspace in many different angles which can easily disrupt the AR image tracking. To solve this problem, multiple AR tag tracking allows the worker to look at an object from different perspectives without losing tracking. Large sheets of textured tags can then be placed and calibrated to show the hands in a particular point in space.

Since the hands are 3D objects, a shadow can be generated from these 3D hands. The utility of the shadow helps ground hands to the real world. Otherwise the hands appear to be

floating without point of reference as to how high they are off the table. The AR tag has a flat shadow plane object that is parallel to the surface of the table. Hence shadows generated on this plan appear to be on the table, giving the 3D hands a sense of height, even without a stereoscopic display.

Head Mounted Display

Figure 4. A worker wearing a Google Cardboard head mounted display working with two hands on a task.

A modified Google Cardboard was used as a stereoscopic head mounted display. I had originally considered using a see-through head mounted display, however I have since switched to a modified Google Cardboard. While the see-through glasses technology is unique, there is no need to use the see-through capabilities of this type of head mounted display. Since the digital display itself is always on in a see-through, the focus is primarily on the display as the virtual hands need to be seen by the working user. This also complicates the analysis as users need to gaze switch between seeing through the glasses and seeing the gestures. Instead, with Google Cardboard a more powerful device (i.e, smartphone) can be used to render the 3D hands in stereoscopic 3D and ignore the effects of gaze switching using a see through head mounted display.

Since the mobile phone only has one camera, only a 2D video feed can be rendered in the background. However the hand capture data is 3D which can be rendered in with stereoscopic vision using Google Cardboard which can be enabled arbitrarily. Here we can validate weather a 2D background is sufficient for enabling users to utilize depth cues in their gestures.

PILOT STUDY

To address whether 3D hand captures in a mobile video conferencing environment is more effective than hand shadows, I designed a within subjects pilot study to evaluate the prototype system. The goal is to evaluate and determine design factors so that future work can look into the effectiveness of hand gestures under various rendering scenarios.

Tasks

The task I have designed was modelled after a construction task. Construction tasks in remote collaboration generally have pieces in which a working user uses to build on an existing structure or create an entirely new structure. A remote partner is called in to assist the work in construction as the helper may have additional information or specialization that the worker needs to complete the task. With an existing structure, users will have to analyze what has already been done and use their problem solving skills to determine which pieces to use to complete the object. A construction task is ideal for evaluating how effective hand gestures are at disambiguating communication since users will encounter many different parts and pieces to analyze. Especially in a 3D task where pieces may need to be oriented and positioned exactly.

Pairs of participants are placed in the same room and are given a randomized list of 4 instructions. They are not allowed to look at each other but may communicate by voice and through the system. The instructions involved taking cards and dice with ambiguous symbols and placing them in cells. Participants must place the pieces accurately down to both the position and orientation. The goal is to complete all 4 instructions within 7 minutes.

Users are given either a helper or worker role throughout the entire study. The worker wears the head mounted display, operates on an immovable structure, and receives the gestures from the helper. The helper is given the instructions to follow and makes gestures to communicate those instructions to the worker. A brief training task was assigned to familiarize the system and the instructions with the participants.

Instructions were designed with the intention of making users look around to analyze the structure. Each instruction card is a picture clue showing the exact location and orientation to place one of the two types of pieces. Since any of the four inner walls are marked once, workers need to look at the marked walls at various angles to find the exact location. This effectively forces users to work at different angles so that we can examine the effects of camera work on gesturing.

Once the location has been determined, the pair needs to find the correct piece to place at that location. There are two different dice sets who only differ by 3 symbols as well as 6 different cards. Each symbol is a wing ding to make it difficult to verbally describe so that users will need to make gestures to communicate more efficiently. Furthermore, because orientation is important to completing each instruction, helpers may also need to use gestures to communicate direction and amount of turning.

Study Design

To evaluate this system against prior work involving 2D hand gestures, I designed a within subjects study to compare hand shadows against the 3D hands and the use of stereoscopic displays. The following is a table of the configurations:

	Shadows Only	Hand & Shadow
2D display	A	B
Stereoscopic	-	C

Table 1. Configurations used for the within subjects study.

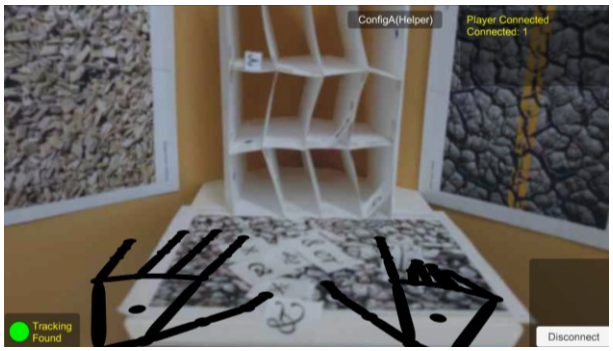

Figure 5. Configuration A, hand shadows only as seen by the helper's view.

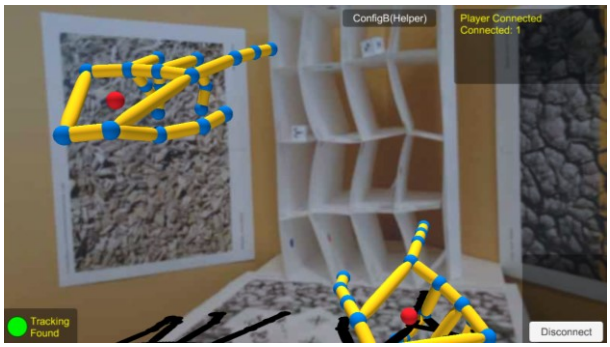

Figure 6. Configuration B, 3D hands and shadow as seen by the helper's view at an angle.

Configuration A. Shadows only, participants are only given a hand shadow. This is similar to that of prior work [e.g. 5,7,8,9,10]. Augmented reality binds these shadows along the work table.

Configuration B. Hands with Shadow, a 3D skeletal hand model along with a shadow. The addition of the hand adds another dimension of gestural freedom. This allows users to easily make gestures in the 3D mobile workspace.

Configuration C. Stereoscopic vision with hands and shadow, is exactly like configuration B but with stereoscopic vision enabled. This allows the worker to perceive depth cues which communicates scale.

Stereoscopic vision with shadows only was taken out as it is nearly equivalent to having shadows only. The utility of stereoscopic shadows is limited as they are always bound to the same plane that the AR tag provides. Since they are always bound to this plane, the shadows will always appear to be at the same height in the real world making forward and backwards movements ambiguous when referring to objects that the shadow cannot reach.

A brief interview was conducted between each configuration and after the study. Feedback for the system was recorded manually to determine design factors for a higher fidelity prototype and study.

FINDINGS

A total of 3 pairs of participants took part in the pilot study, 1 of which did not complete all configurations due to major technical issues but provided valuable feedback. All pairs of participants were familiar with each other.

N=2	Shadow only	Hands & Shadow	Stereoscopic Hands & Shadow
Time (s)	330	360	247
Preference	0%	0%	100%
Errors	0	1	0

Table 2. Quantitative data across pairs of participants. Only two pairs of participants completed all configurations.

Poor hand capture fidelity. Helpers were unable to communicate their gestures most of the time as their hands were moving too rapidly and the postures were too complex for the capture camera. This was a source of frustration as participants needed to learn how to slow down their gestures. Non open palm gestures such as grasping and loosely pointing fingers were not well recognized by the camera. Eventually the helper develops workaround hand gestures that the pair mutually agree on to mean particular things. An example of a workaround was making long waving motions to indicate how much more turning one needs to make on a dice. Despite the poor capture fidelity, helpers continued to make hand gestures anyways but focused on verbally describing their commands as they did not trust the system to capture all of their gestures.

Small and unclear interaction volume. Participants understood the concept of the interaction volume that was configured to the space front of the monitor; however this volume was used in many unintentional ways. Most participants understood what the space was for and deliberately made gestures outside of the volume to help themselves think. An example of this had one helper make a circle gesture to themselves to make sure that the direction they were referring to made sense when they communicated the instruction over the screen.

Participants also felt that the size of the interaction volume was small because the monitor was obstructing their ability to move their hands towards the screen. If the worker moved their camera close enough to the structure, only a small portion of the 3D hands would show up in the worker's view. Because the augmented reality system anchors the hands in a fixed position in the real world, helpers needed to move their hands further towards the monitor to keep their virtual

hand in view for the worker. However the monitor obstructs their ability to keep that virtual hand in view.

Poor video capture resolution. All workers initially needed some time to adjust to the HMD's view of the world through the digital camera. Workers found it difficult to coordinate their hands due to the 2D capture of the real world and the novelty of the HMD.

However what hindered their ability to complete tasks efficiently was the low resolution camera. Workers often could not differentiate the symbols on the cards and dice. Instead they relied on the helper to make the distinction. Even then, workers had to bring the piece very close to the camera in order to assist the helper with making the distinction.

Verbal instructions are more preferred. In a post study interview, participants agreed that verbal instruction were preferred over using gestures. A large part was due to the distrust with the system and since it felt more efficient to verbally communicate than to figure out why the gesture was not being captured. When asked between each configuration to reevaluate their strategy, only a more restricted codified gesture was suggested or simply not using gestures at all.

Configuration C was the most preferred. Stereoscopic hands and shadow was the most preferred configuration by all participants. However when questioned why, participants felt that the configuration was too similar to Configuration B: Hands and shadow. Hence they choose configuration C because it was the most featureful of the three.

Limitations

Since there were only two pairs of participants in the study, it is difficult to draw any valid conclusions from this data. While it may appear that the completion time for configuration C is the lowest, there are not enough participants to conclude that configuration C is the best in terms of speed. Likewise for any other conclusion drawn. Instead I will focus my analysis on the qualitative feedback I had acquired from the studies.

FUTURE WORK

While my findings are overall negative, the utility of hand gestures still needs to be investigated. Many of the criticisms arose due to the technical limitations of the system rather than a limitation of human gesturing itself.

Higher fidelity hand capture system. A higher fidelity hand capture system should be addressed to allow helpers to make gestures freely without fear of losing tracking. Such a system should be able to capture rapid hand movement and complex hand postures regardless of the angle made. Establishing trust with the system should allow helpers to make such gestures freely and thus we may see a greater variety of gestures used that we may not have encountered otherwise. Which opens the possibilities to studies involving hand technique tutorials and physical props.

Small and unclear interaction volume. An inadequately sized interaction volume prevents users from making their gestures in the way they intended. This can be improved by increasing the volume in which the hands can be captured, and preventing the monitor from becoming an obstacle. In addition, the scale of which the capture needs to be aligned with the captured world as well. Participants also felt the misalignment felt odd as their gesturing was not one-to one. This is important as helpers may need to replicate gestures to scale where precision is needed. Such a system can potentially enable a helper to grasp the scale of real world objects simply by comparing the distances between their hands. Additionally, the interaction volume should also support self-gesturing to allow helpers to think to themselves.

Stereoscopic vision with stereo cameras. Currently, the system renders a 2D background as the video feed, but the hands are displayed in stereoscopic 3D. Workers had to take time to adjust to the lack of depth information of the real world. It would be helpful to utilize a set of stereo cameras instead to capture the world. If the hands are correctly calibrated to appear correctly to the scale of the world, then we may be able to examine the benefits of having stereoscopic vision on gestures. In such a system, gestures could be perceived in exact scale relative to the real world. Then the question is whether a helping user may need other aids to help make these gestures in the exact scale.

Improving task design. One participant mentioned that the gesturing required for the tasks were too simplistic and involved only pointing. The instructions were originally designed to examine the efficiency of hand gestures which did not incorporate more complex gestures. Future task designs should consider how gestures can potentially be used to communicate more information at once. Information such as temporal cues (i.e. rhythm and timing) or using particular a particular hand can be examined under different conditions such as camera mobility, AR stabilization, and stereoscopic vision.

Complexity of structure. The structure was designed so that the AR tags are always in view regardless of what angle a worker looks. But this is limited because this does not involve moving around the structure or conditional instructions. The tasks are largely information driven by the helper where the worker simply complies to the instructions given by the helper. However communication is rarely unidirectional and hence tasks should also involve communicating information about the workspace.

Other minor technical issues. There are a number of other minor technical issues that contribute as design factors such as poor video quality and comfort of the HMD. A higher quality video feed in terms of resolution should be used to help participants focus on completing the task. The head mounted display should be comfortable for long periods of time and the study design should allow for breaks in-between.

CONCLUSION

When designing mobile video conferencing tools for collaboration with hand gestures, a number of key design factors need to be considered. Prior work have looked at fixed camera, 2D gesturing tools and found that gesturing is a rich communication tool for collaboration. However few have adequately addressed the external validity to real world tasks involving mobile video cameras. Which adds design challenges when supporting scenarios involving 3D objects at multiple and moving angles. My system aims to explore the use of gesturing for collaboration using AR techniques to support 3D tasks.

Using a depth capture of hands, my system builds on prior work by allowing virtual hand gestures to be made in 3D which attempts to resolve perspective issues when working with a mobile camera. In addition, my system integrates AR stabilization to hand gestures which aids users in making and maintaining gestures in their intended contexts since video feeds can be unstable in mobile scenarios.

My work contributes several potential design factors when designing collaborative gesture systems for remote mobile assistance. Higher fidelity hand capture system, interaction volume configuration heuristics, and task design are factors that need to be addressed in future exploration. Further investigation needs to be done to explore the use of other types of gestures for other tasks beyond construction type tasks. Other tasks may have other design constraints that affect the presentation of a gesture and their abilities to express more information than verbal communication. Whether or not hand gestures are well supported by the mobility of the camera, there is greater potential for exploring gesture support in remote, mobile 3D environments.

REFERENCES

1. Bill Buxton. 2009. Mediaspace – Meaningspace – Meetingspace. *In Media Space: 20+ Years of Mediated Life*, S. Harrison (Ed.). Springer, London, UK, 217-231.
2. Steffen Gauglitz, Cha Lee, Matthew Turk, and Tobias Höllerer. 2012. Integrating the physical environment into mobile remote collaboration. In Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services (MobileHCI '12). ACM, New York, NY, USA, 241-250. DOI=10.1145/2371574.2371610 http://doi.acm.org/10.1145/2371574.2371610
3. Weidong Huang, Leila Alem, and Jalal Albasri. 2011. HandsInAir: A Wearable System for Remote Collaboration. *CoRR, abs/1112.1742.* http://arxiv.org/abs/1112.1742
4. Brennan Jones, Anna Witcraft, Scott Bateman, Carman Neustaedter, and Anthony Tang. 2015. Mechanics of Camera Work in Mobile Video Collaboration. *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 957-966. DOI=10.1145/2702123.2702345 http://doi.acm.org/10.1145/2702123.2702345
5. Aaron M. Genest, Carl Gutwin, Anthony Tang, Michael Kalyn, and Zenja Ivkovic. 2013. KinectArms: a toolkit for capturing and displaying arm embodiments in distributed tabletop groupware. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13). ACM, New York, NY, USA, 157-166. DOI=10.1145/2441776.2441796 http://doi.acm.org/10.1145/2441776.2441796
6. Seungwon Kim, Gun Lee, Nobuchika Sakata, Mark Billinghurst, "Improving co-presence with augmented visual communication cues for sharing experience through video conference", *ISMAR, 2014, 2014 IEEE International Symposium on Mixed and Augmented Reality* (ISMAR*)*, pp. 83-92, doi:10.1109/ISMAR.2014.6948412
7. David Kirk, Andy Crabtree, and Tom Rodden. Ways of the Hands. in *Proceedings of the 9th European Conference on Computer-Supported Cooperative Work (Paris, France, 2005),* Springer, 1-21.
8. Anthony Tang, Michel Pahud, Kori Inkpen, Hrvoje Benko, John C. Tang, and Bill Buxton. 2010. Three's company: understanding communication channels in three-way distributed collaboration. *In Proceedings of the 2010 ACM conference on Computer supported cooperative work* (CSCW '10). ACM, New York, NY, USA, 271-280. DOI=10.1145/1718918.1718969 http://doi.acm.org/10.1145/1718918.1718969
9. John C. Tang and Scott Minneman. 1991. VideoWhiteboard: video shadows to support remote collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '91), Scott P. Robertson, Gary M. Olson, and Judith S. Olson (Eds.). ACM, New York, NY, USA, 315-322. DOI=http://dx.doi.org/10.1145/108844.108932
10. John C. Tang and Scott L. Minneman. 1991. Videodraw: a video interface for collaborative drawing. *ACM Trans. Inf. Syst. 9, 2* (April 1991), 170-184. DOI=http://dx.doi.org/10.1145/123078.128729
11. Rajinder S. Sodhi, Brett R. Jones, David Forsyth, Brian P. Bailey, and Giuliano Maciocci. 2013. BeThere: 3D mobile collaboration with spatial input. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13). ACM, New York, NY, USA, 179-188. DOI=10.1145/2470654.2470679 http://doi.acm.org/10.1145/2470654.2470679