Object-Based Touch Manipulation for Remote Guidance of Physical Tasks

Matt Adcock[*†‡] **Dulitha Ranatunga**[*†] **Ross Smith, Bruce H. Thomas**[‡]

[*]CSIRO
Canberra, Australia

[†]Australian National University
Canberra, Australia

[‡]University of South Australia
Mawson Lakes, Australia

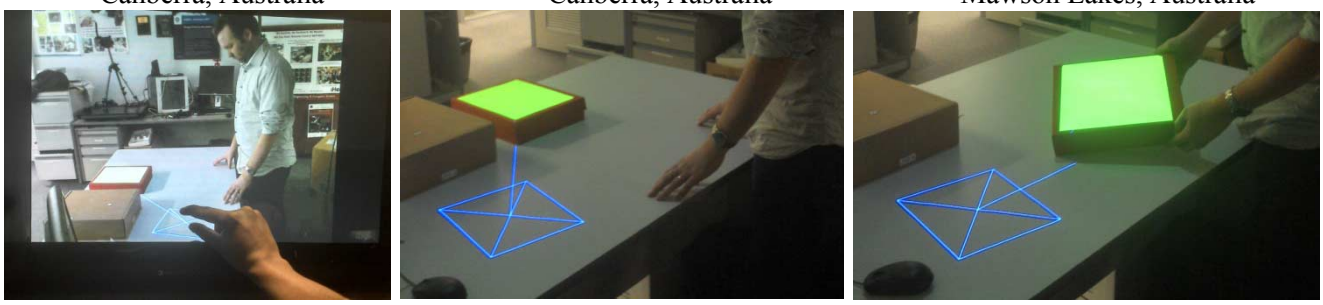

Figure 1. (left) An expert directly specifies a rotation and translation of the box using multi-touch. (center) Rotated and translated visual cue from workers' perspective. (right) As the worker moves the box, the SAR cue is updated.

ABSTRACT

This paper presents a spatial multi-touch system for the remote guidance of physical tasks that uses semantic information about the physical properties of the environment. It enables a remote expert to observe a video feed of the local worker's environment and directly specify object movements via a touch display. Visual feedback for the gestures is displayed directly in the local worker's physical environment with Spatial Augmented Reality and observed by the remote expert through the video feed. A virtual representation of the physical environment is captured with a Kinect that facilitates the context-based interactions. We evaluate two methods of remote worker interaction, object-based and sketch-based, and also investigate the impact of two camera positions, top and side, for task performance. Our results indicate translation and aggregate tasks could be more accurately performed via the object based technique when the top-down camera feed was used. While, in the case of the side on camera view, sketching was faster and rotations were more accurate. We also found that for object-based interactions the top view was better on all four of our measured criteria, while for sketching no significant difference was found between camera views.

Categories and Subject Descriptors

H.5.1. [Information interfaces and presentation] Artificial, augmented and virtual realities; H.5.2. [Information interfaces and presentation] Input devices and strategies.

General Terms

Human Factors, Design, Experimentation.

Keywords

Spatially Augmented Reality; Remote Guidance; Object Manipulation; Multi touch interaction; 3D CHI.

SUI'14, October 4–5, 2014, Honolulu, HI, USA.

http://dx.doi.org/10.1145/2659766.2659768

INTRODUCTION

Remote guidance takes place when multiple participants in different locations work together to perform a task involving physical objects. Typically the scenario involves an 'expert' with specialized knowledge in the task situated remotely from the physical task environment. They collaborate, using available communication tools, with a 'worker' who is co-located and able to physically interact with the task environment.

In our research, one of the task scenarios we aim to support is a science laboratory technician (worker) at a lab-bench, receiving guidance from a remote supervising scientist (expert). Sometimes that scientist can be remote due to geography, and other times there might be a quarantine barrier in place that is costly to cross[1]. We are therefore motivated to develop systems that support remote guidance of part-tasks such as spatial arrangement (of equipment and specimens) on the research bench top.

There are many different approaches to improving the communication tools and capabilities of the remote helper to enable efficient guiding techniques. These include video-conference situations [16], view sharing based head mounted display (HMD) systems [15] and through the use of spatial augmented reality. Unfortunately, in video-conference based systems, the expert is only able to provide verbal cues in response to the visual feed. HMD systems are a form of augmented reality, however they suffer from being encumbering wearable devices that can limit the helpers' freedom to move or operate.

Spatial Augmented Reality (SAR) systems use projectors to display computer generated graphics directly onto the physical environment [3]. SAR provides the benefit of augmentation without wearable or hand-held devices. Furthermore, since the workspace is augmented spatially, the same graphics can potentially be seen by multiple workers using naked-eye stereo affordances (i.e. simply looking at real-world objects).

Some SAR based remote guidance systems have superimposed the expert's sketches or hand gestures into the worker's

[1] This example is inspired though working on the CSIRO's collaboration platform for the Australian Animal Health Laboratory to facilitate the management of exotic diseases.

environment, but they have typically not used any semantic information about the 3D properties of the physical objects.

We propose a method that encompasses the benefits of SAR and incorporates interaction techniques previously developed for 2D manipulation of virtual 3D objects [18]. As shown in Figure 1 and Figure 2, users are able to touch objects on a video feed and manipulate a hidden virtual representation of the physical work environment. Changes made in the virtual scene are used to generate augmentations in the real world that act as spatially aligned guidance for the worker. As objects are moved within the physical scene, the virtual model is updated and the augmentations can also be updated accordingly.

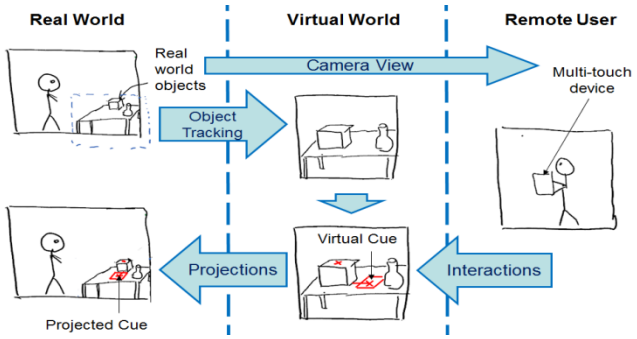

Figure 2. A workflow of how a virtual scene can be used to mediate in SAR-based remote guidance

The main contributions of this paper are as follows:

- We extend the concept of object based touch manipulation originally introduced in a works-in-progress paper by Ranatunga et al. [25]
- We have implemented a prototype system that allows an expert to use multi-touch gestures to translate, rotate and annotate an object on a video feed, via a proxy target, causing meaningful projections onto the real world.
- A user study showed that object manipulation is able to improve upon sketching based SAR remote guidance from a view point perpendicular to the work surface.

In the following sections we review some related work. We then review the concept of object-based touch manipulation, including a focus on how some interaction techniques can be employed. We then present an evaluation of two such techniques using our prototype system under two possible camera configurations.

RELATED WORK

Our approach draws from a variety of fields, including remote guidance, SAR guidance, 3D multi-touch and proxy-based 2D image manipulation.

Many previous remote guidance systems involve a shared video feed of the worker's environment, composited together with the expert's hands [16], pinpoints [7], or sketch annotations [6,23]. A comparison of existing remote gesture technologies was conducted by Kirk and Stanton Fraser [14]. In particular, they evaluated two SAR-like methods: projected hands, and projected sketches. In the first setup, the helper's hands were recorded from a top down camera and the video feed was directly projected onto the surface of the worker's location. The second approach extended the first by including a digital whiteboard that allowed for the combination of sketching as well as the hands. In both systems, visual feedback to the expert was via a camera feed of the augmented physical world shared with the worker. There was no model of the scene within the system so it was up to the experts to orient themselves, define reference points, and use relative movements.

In some non-collaborative scenarios, SAR has been shown to benefit guidance of physical tasks. LightGuide is a system that projected guidance cues onto a moving hand to follow a particular path [29]. Rosenthal et al. used micro-projectors to complement on-screen instructions for manual task guidance and determined situations in which guidance improved and decreased performance tasks [28]. Marner et al. [19] showed that, for a procedural button pushing task, SAR overlays led to faster task completion speeds and fewer errors.

Tsimeris explored a variety of projected cues that could assist in the arrangement of physical objects [33]. While the system did incorporate a 2D GUI authoring tool, capable of real-time manipulation, it did not facilitate a collaborative remote guidance scenario, nor any video feedback to a remote expert.

Henderson has shown that dynamic instructions overlaid on or near objects is preferred for psychomotor tasks and is significantly more efficient when compared to a nearby LCD display [11]. That study was specific to HMD based augmented reality and the guidance was entirely pre-authored.

Tecchia et al. employed depth cameras to stream real time virtual representations of the workspace for remote guidance [32]. The system required the expert to wear an HMD showing a depth camera feed from the workspace, while the worker watched a combination of that feed and a depth camera feed taken of the expert's hands. The BeThere system [30] used a similar techniques but required the depth camera and feedback display to be held in the non-dominant hand.

The RemoteFusion system [1] also used depth cameras to capture a 3D scene for a remote expert. The expert could draw on the 3D model and the sketches were projected using SAR onto the physical workspace. Again, unlike the system we present in this paper, RemoteFusion did not perform any segmentation or tracking of the individual objects that made up the scene.

TeleAdvisor [9] was a SAR based remote guidance system that projected a green square into the scene and used it to estimate the location of roughly planar surfaces. The Sticky Light system [8] similarly tracked planar surfaces (or almost any other object) using fiducial markers. Although the expert's annotations 'stuck' to the tracked objects, the input method was still just 'sketching'.

Suenaga et al. [31] implemented a tele-instruction system which used a shared AR space. The expert would manipulate a single ultrasonic probe via a visualization named a 'Web-Mark' which was projected onto the body of a patient. Hiura [12] developed a tele-direction system that creates a virtual model of objects to allow for projected annotations. That system operated somewhat like a 3D version of the copy/paste feature of Wellner and Freeman's Double Digital Desk [34]. We extend these ideas by allowing the expert to manipulate multiple object targets using direct and indirect touch input. Furthermore, our system does not require the expert to first trace the outline of the object(s).

A number of research efforts have explored the ways 2D multi-touch can be used to interact with 3D virtual environments and Liu et al. [18] provide a good comparison of current methods. This is a relatively new field, but already we are seeing useful abstractions emerge which permit different ways of interacting, depending on the required task.

In terms of interaction with captured 2D scenes, there is also related work in video and photo manipulation. Dragicevic et al. [5] described a system for direct manipulation of video playback using inter-frame optical flow. Proxy-based manipulation of photographed objects has been presented by Zheng et al. [35] and Chen et al. [4] using, respectively, cuboid and cylinder proxies. More recently, this idea was extended by Kholgade et al. [13] with more complex 3D proxies sourced from online libraries. In our new system we take inspiration from these approaches, and use proxies to facilitate a remote expert's touch interaction with a live video stream.

OBJECT BASED REMOTE GUIDANCE

We extended the concept of object based touch manipulation for remote guidance from Ranatunga et al. [25] with spatial direct and indirect touch capability and a sketching feature. Object based touch manipulation for remote guidance applies 3D spatial knowledge of a physical workspace to allow a remote expert to denote understandable guidance via a multi-touch interface.

Interacting with 3D information through a 2D interface is an inherently difficult problem [26]. Object based touch manipulation for remote guidance constrains the expert's interactions to actions that make sense in the physical 3D environment, such as laboratory equipment remaining in contact with a workbench, and aims to make the presented information easier to specify and more understandable. To enable a mapping from the remote expert's interaction with a 2D multi-touch screen into the 3D physical workspace, a virtual world is constructed and maintained to mirror the real world (see Figure 3). Proxy objects in the virtual world can then be interacted with through the use of multi-touch gestures. As soon as the location or orientation of a proxy differs from its respective physical object, SAR visual cues are generated to assist the worker in manipulating the physical object to match the according target pose. The proxy is effectively used by the remote expert to specify the 'end goal' of a physical manipulation for the worker. A more detailed description of object based touch manipulation for remote guidance is presented in Ranatunga et al. [25].

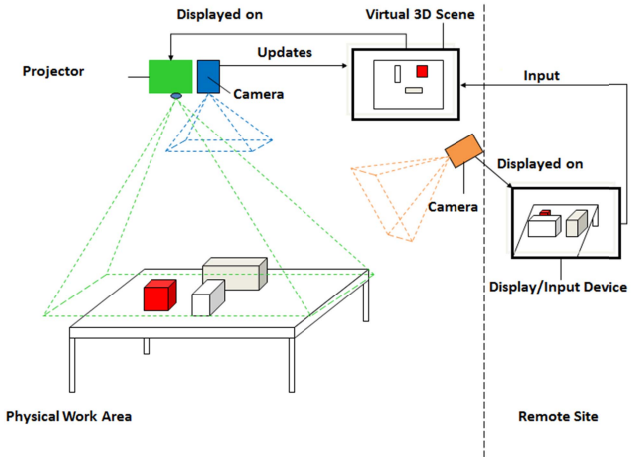

Figure 3. An overview of our system concept

SAR systems typically only permit projection of graphics onto the surfaces of physical objects. In sketching based SAR systems, it may be hard for the helper to visualize a meaningful 2D projection for some manipulations (such as rotations) and from some viewing angles (such as anything other than directly overhead). The proposed system deals with this issue by automatically generating the visualizations.

Spatial Direct and Indirect Touch

Our system allows 2D gestures performed on a multi-touch display to select and translate based on the constraints of the physical environment. Used by the remote expert, only gestures and a video feed are shown on the display (as shown in Figure 4). The visual cues are presented to the local worker through SAR in the 3D physical space (which are fed back to the remote expert automatically in the captured video). We describe this translation as "spatial direct touch" manipulation [20] as the projected target remains under the user's finger no matter where it moves.

A second finger allows for indirect rotation by moving up and down the vertical axis of the screen (see Figure 5). The rotation finger can be used to rotate the object from anywhere on the screen and both fingers can be moved independently, allowing for simultaneous rotation and translation.

An alternative to the direct-translation/indirect-rotation technique could have been to aim for entirely direct manipulation. However, as we are essentially manipulating real physical objects (albeit via virtual proxy), there are physical constraints to consider.

One technique, now widespread in photo viewing apps on multi-touch tablets and phones, is drag to translate, pinch to scale and twist to rotate [27]. In this case, there is no useful analogue to 'scale' in the physical context, especially when we assume that boxes cannot float in mid air above the table. Attempting to use this technique therefore results in inconsistencies such as can be seen in Figure 6.

We note here that Liu et al. [18] have previously demonstrated that a non-direct manipulation method can out-perform a direct manipulation method in some contexts. Others have also reported that separating the degrees of freedom in the interface can improve the accuracy because they do not require the user to modify one aspect in order to refine another [21,22].

Figure 4. (Left) Selecting an object, such as one of the boxes, by touching it with one finger on the touch screen, selects it and creates a virtual target. The target is projected directly onto the physical table. (Right) Moving the finger on the touch screen causes the projected target to move accordingly and appear to stay directly under the user's finger.

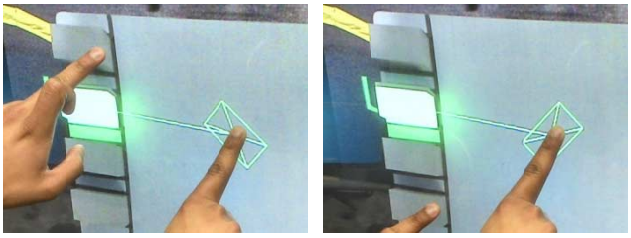

Figure 5. Using a second finger to swipe up and down anywhere on the touch screen will cause the projected target to rotate around the first finger's current position.

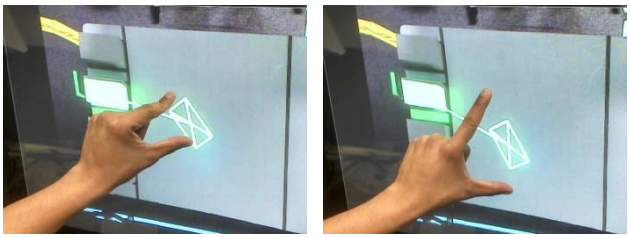

Figure 6. An alternative interaction technique that uses two fingers to drag/pinch/rotate can experience inconsistencies due to the inability of physical objects to scale up or down in the real world.

Sketching

We expect that, for some tasks, sketch based remote guidance will still be useful and therefore we have aimed for our object-based implementation to be compatible with sketching. In our prototype, sketching is implemented in a similar way to the Sticky Light system described by Gunn and Adcock [8], in which geometric knowledge of the physical scene is used to ensure sketched annotations are projected in the correct locations.

When the sketching mode is active, users can use a single finger as input. To avoid finger occlusion issues, a screen-space offset is applied (in a similar way to the *take-off* technique by Potter et al. [24]), giving the impression that the 'digital ink' is flowing from the very tip of the user's finger. Their touch points are sensed and sent to the hidden virtual environment clone of the workspace. There, the touch points are ray-cast onto the respective virtual object and a glowing virtual line is created. This results in the glowing lines being displayed by the projector in the correct location on the respective physical objects.

Figure 7. Sketching with a finger on the remote expert's touch screen results in glowing lines projected onto the physical objects in the workspace.

Prototype Implementation

A prototype system was constructed to demonstrate and evaluate the concept of object based touch manipulation for remote guidance, and the design of this system was inspired by a lab workbench context. The system was implemented with an overhead Kinect depth camera, a projector and an external 'side-on' camera (Lanir et al. [17] found an optimal camera viewpoint can vary with the task). The remote expert interacts with a multi-touch screen. Since the virtual world is hidden from the expert, the visual feedback to the expert is, in fact, the same as the augmented environment provided to the worker. This enforces a strict shared view of the workplace. An example of this can be seen in Figure 1. This overall system arrangement can be seen in Figure 3 and Figure 8 and is described further in [25].

EVALUATION

We designed and conducted a user study to investigate the performance of this new remote guidance system. The user study was conducted as a counterbalanced 2x2 within-subjects design, with the independent variables being two camera angles that could be practically installed in a science lab (**top**-down or from-the-**side**) and type of SAR interaction (our new **object**-based manipulation or 'traditional' **sketch**ing). There were 12 pairs of participants with each pair working together to complete tasks from each of the 4 conditions:

top_object – object based input on the top camera view
top_sketch – sketch based input on the top camera view
side_object – object based input on the side camera view
side_sketch – sketch based input on the side camera view

This study consisted 19 males and 5 females. Participants reported they were generally not familiar with augmented reality, SAR or remote guidance, but most were very familiar with touch screens.

Setup

Within each pair, participants were randomly assigned a role (either worker or expert), which they kept for the entire duration of the trial i.e. the worker and expert did not swap roles. Both the worker and expert were situated in the same room and could talk naturally with each other. During each trial, the worker was unable to see the expert due to the erection of a physical barrier to block visibility. The room was internal to the building and lighting was kept consistent through the use of indoor lighting. The experimenter was situated with visibility to both environments, giving the ability to take notes without moving around in a distracting manner. The expert had two main displays; the touch screen with which they interacted and a goal screen which displayed instructions.

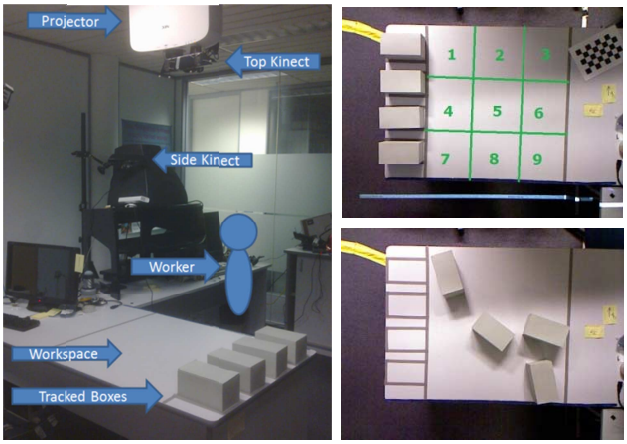

Figure 8. (Left) The workspace used in the evaluation (Top Right) The grid pattern used to design the arrangement task instructions. (Bottom Right) An example arrangement for the remote expert - using grid squares 1, 5, 6 and 9.

Tasks

Each pair of participants would carry out 20 trials in total. Within each trial, participants were tasked with spatially arranging the four identical boxes shown in Figure 8 into positions on a table. It did not matter which box went where, but the worker would not previously know where any of the boxes were supposed to go. Instead, the expert was shown an image such as Figure 8 (bottom right) on an auxiliary computer monitor, and was told to convey the positions to the worker with the aid of object manipulation or sketching. This task is designed to be similar to the spatial arrangement of equipment on a lab-bench. In order to control

order effects, a 4x4 balanced Latin Square was used to determine the order that the four conditions were administered.

To maintain a roughly consistent difficulty across each trial, 20 task goal images were generated according to a specific grid location as highlighted in Figure 8 (top right). The angular rotation of each of the 4 boxes was randomly generated such that the total sum of all the rotation would be 180° so as to maintain roughly equal rotational difficulty. The 20 task goals were *clustered* according to their use of similar grid positions such that they could be mapped in structured random order to the 20 trials of each pair. This structured mapping meant that the practice run of each condition would be from Cluster A, the second trial from B, and so on until the last trial, which was always from Cluster E.

Experimental Procedure

The study was conducted with one pair of participants at a time. The participants were first informed of their randomly assigned role (worker or expert). The experimenter then provided a demonstration of the system from both the expert's perspective and the worker's perspective, with both participants seeing both demonstrations. The demonstration also introduced a user study management application including how to stop and start each trial.

Then, for each condition in the experiment:

(a) The experimenter would demonstrate how to interact with the respective particular condition.

(b) There would be one practice trial

(c) There would be 4 recorded trials

(d) In each trial the following procedure took place:

- The user study management application would light up the starting positions for boxes and wait for the worker to confirm they are reset.
- When both the worker and expert were ready, the expert could press a key to start the timer and the trial.
- The new task goal picture would appear on the experts' second screen.
- The pair would communicate to perform the task, while the experimenter took down any notes and observations.
- The expert would press a keyboard key to stop the timer and end the trial, causing the system to record their box arrangement.

After all the trials were over, participants were then asked to complete a survey to gather their opinions of the tasks they had undertaken.

Data Collection

The quantities measured during each trial are detailed in Table 1. *Time* is used as the proxy for task efficiency, and *Cost* is inversely proportional to task accuracy.

In addition to these measurements, we also kept the pose (position and orientation) of each box at the start and end of each trial such that scores based off new metrics could be calculated. A screen shot from both cameras was also taken at the end of each trial. These screen shots were used in the case of outliers during the results analysis to visually verify whether any errors occurred in the systems' gathering of data. All of the data was written to a few different files, tagged by a unique identifier associated with the participant ID and time. After all of the trials were complete, a separate python script was written that allowed for easy compilation of data into a single .csv file for statistical analysis.

Upon inspection of the data collected by the system, there were a total of 7 values that had incorrect data recorded. These values had unusually high costs for the translation, rotation and aggregate cost measures. Each value was compared with the screen shots that were taken at the end of each trial. By visual analysis and comparison of the pose matrices of each box, it was determined that the expert had ended the trial while the worker was occluding a box from the tracking camera's view point. Since this invalidated the data from those trials, those 7 values were removed and replaced with 'missing' values in the statistics software tool which, in turn, compensated for this in the analysis.

Measure	Recording/Calculation Method
Total Time (ms)	The difference between the time of the trial starting and the time of the touch screen application being told the trial was over. (Error: 1ms)
Translation Cost per box (mm)	The tracking system reported the 4x4 pose matrix of each box. The target pose was saved when the tests were generated, and we record the Euclidean distance between the center of the target and the result box.
Rotation Cost per box (radians)	Given a target pose matrix A, and end pose matrix B, we calculate the transformation: C = inverse(A)*B. C represents the transformation to get from one box to the other. We then decompose the rotation part of C into a vector and angle component. The cost is the angle in radians.
Aggregate Cost per box (mm)	The Euclidean distance between the top four corners of the target box (defined by depth values closest to the tracking camera) and the top four corners of the result box.
Total Translation Cost (mm)	The sum total of the translation cost per box. (Error: <11mm)
Total Rotation Cost (radians)	The sum total of the rotation cost per box. (Error: <0.1)
Total Aggregate Cost (mm)	The sum total of the aggregate cost per box. (Error: 68mm)

Table 1. Quantitative Data Recorded

Measurement Error

The 3D tracking system was used for measuring the translation cost, rotation cost and aggregate cost. The tracking system data; however, had a certain level of noise in its accuracy. To calculate the magnitude of this error before conducting the user study, we captured the tracking data of 4 stationary boxes multiple times and measured the range of values. The maximum error of the total translation cost was found to be 10.4mm (mean of 2.6mm per box). The maximum error of the total rotation cost was found to be 0.06 radians (< 1 degree per box). The maximum error of the total aggregate cost was 67.1mm (mean of 16.8mm per box). The time measurement was meant to be accurate to the nanosecond level; however, the signal to start and stop the timer is also affected by network latency. In either case, this (<1ms, estimated) error would be consistent between all trials and is not formally tested. These values are included in Table 1.

RESULTS

In this section, we report the results and analysis of each of the four main quantitative measures: Task Time, Translation Cost, Rotation Cost and Aggregate Cost. Note that these are summarized in Table 2 at the start of the next section. We also report on some results from the user questionnaire.

Time

Side_object had the longest trial time with a mean of 117.7 seconds and a standard error of 6.90 seconds. The condition with the shortest average completion time was **top_object** with a mean of 91.2 seconds and a standard error of 5.42 seconds. The overall mean was 106.27 seconds.

A split plot analysis of variance was conducted on the data after a natural log transform was applied. This transform was used to satisfy the assumption of homogeneity in the data.

The spit plot used the four conditions as the main variable and the split on the test cluster whilst adjusting for the effects of different pairs and orders. When tested for differences in completion time, this analysis found a statistical difference between the four conditions ($F(3,24)=3.79$, $p<0.05$). The analysis also showed there was no significant difference in time between test clusters ($F(3,132)=1.60$, $p>0.05$) and even less significance between the different conditions within test clusters ($F(9,132)=0.87$, $p>0.05$).

Performing the Fishers least significant difference (LSD) post-hoc test revealed that **side_object** took significantly longer (at the 5% level) than the other conditions, but the time difference between the other conditions did not differ significantly.

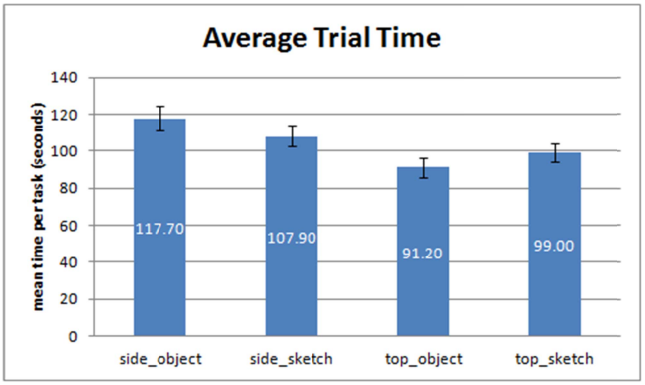

Figure 9. The average completion of each trial by condition. Error bars represent the standard error of the mean, back-transformed from the log transformation.

Cost

Translation Cost

Top_object had the lowest translation cost mean of 122.9mm (SE=7.7mm). The two side conditions had comparable means: **side_object** had 188.4m (SE=11.6mm) and **side_sketch** had 190.2mm (SE=11.7mm).

A split plot analysis of variance was conducted on the (also log transformed) data to test for differences in translation cost between the four conditions. It was found that a significant difference exists ($F(3,24)=8.56$, $p<0.001$) at the condition level, but not between the test clusters ($F(3,125)=0.96$,$p>0.05$).

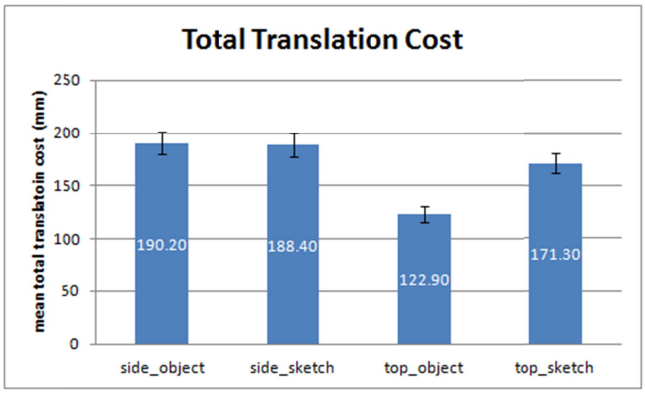

Figure 10. The average total translation error of each trial by condition. Error bars represent the standard error of the mean, back-transformed after a log transformation.

Analysis using Fishers LSD post-hoc test showed that **top_object** performed with a significantly lower cost than the other three conditions. Sketching from the top view is not significantly different under the LSD test at the 5% level in comparison with the side conditions.

Rotation Cost

The total rotation cost is the sum of the angular error in the four boxes of each trial. As an indication of the range of values, **side_object** had highest mean total rotation cost of 1.727 radians (SE = 0.1112 radians) and **side_sketch** had lowest total rotation cost mean of 1.244 radians (SE=0.1112 radians).

A split plot ANOVA was conducted on the untransformed rotation data with the four conditions as the main variable and the split on the test cluster. The ANOVA adjusted for order effects and the differences between pairs. When tested for differences in rotation cost, the analysis found a statistical difference between the four conditions ($F(3,24)=3.58$, $p=0.028$). The same analysis did not show a significant difference between test clusters ($F(3,125)=1.19$, $p=0.315$), however it did reveal a significance difference in the interactions between test cluster and condition ($F(9,125)=2.59$,$p=0.009$).

The Fishers LSD post-hoc test was applied to the conditions to determine that **side_object** had a significantly higher rotation cost compared to the other three conditions. The difference between **side_sketch** and **top_sketch** is not significant at the 5% level.

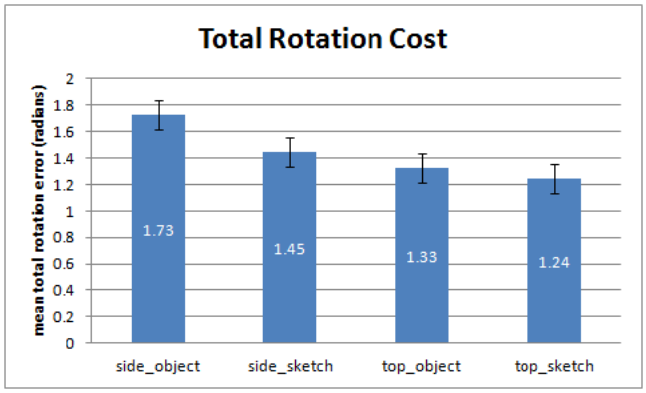

Figure 11. The average total rotation error of each trial by condition for each test cluster . The error bars represent the standard error of the mean.

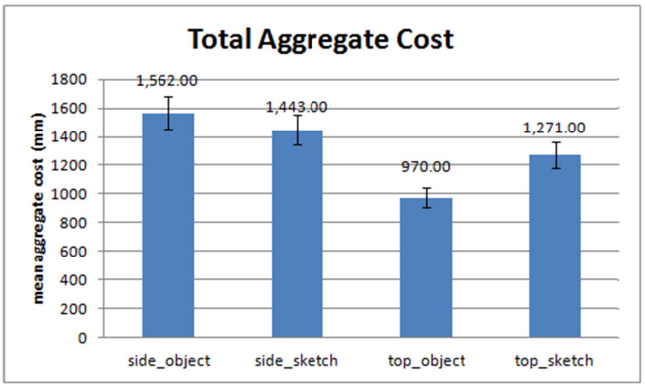

Figure 12. The average total aggregate cost of each trial by condition. Error bars represent the standard error of the mean, back-transformed after a log transformation.

Aggregate Cost

To test for differences in cost among the four conditions a split plot ANOVA was used on the (log transformed) data with adjustments for pair and order effects, and test clusters as the split. This analysis found significant differences between the four conditions ($F(3,24)=7.48, p<0.01$). It also found significant variation at the test cluster level ($F(3,125)=2.83$, $p<0.05$) as well as the interactions between test cluster and condition ($F(9,125)=1.97$, $p<0.05$).

Top_object had the smallest aggregate cost with a mean of 970mm (SE=72.44mm). **Top_sketch** had a mean of 1,443mm (SE=92.45mm), while the **side_object** condition performed with the greatest cost (mean=1,562mm, SE=118.84mm). Fishers' post-hoc LSD test was used to determine the significance of these interactions. It was found that the aggregate cost of **top_object** was significantly lower than the cost of all three other conditions. It was also shown at the 5% level, **top_sketch** had a lesser cost than the aggregate cost of **side_object**, but not a significant enough difference from **side_sketch**.

Qualitative

The participants in the role of 'Expert' were each presented with a series of statements and asked to rate them on a Likert scale. Figure 14(a) shows that **side_sketch** was not considered as "easy to get the hang of", as the other three conditions. All experts agree that **top_object** was "easy to get the hang of". When asked about ease of communication (Figure 14(b)), we see a positive result in **top_sketch**, **top_object** and **side_object**. However, there was no clear indication as to whether or not it is easy to communicate when sketching from the side. The majority of those with an opinion disagreed with the statement. Participants also felt that by the end of the trial, they were proficient at both systems from both angles (Figure 14(c)).

Observations

In general there were key differences in the usage of object and sketch. Each of the pairs independently developed and agreed upon their own strategy for interaction. Many of these strategies repeated themselves between pairs, and this section aims to highlight the different approaches used. Overall, sketching involved a lot less interaction with the system, with a lot of emphasis on the spoken conversation. Generally, the object based trials were conducted with minimal conversation and mostly in silence. This seemed to result in the worker feeling more involved and valuable during sketching.

The interaction graphs in Figure 13 show an example of when the screen was being touched in two corresponding trials, and are indicative of the interaction graphs across most trials. Object manipulation typically involved continual adjustment using the system while the sketching systems tended to be used for initial placements and then verbal adjustments.

Some box layouts were more 'recognizable' than others. For example, three of the boxes in one configuration were in a row, and participants sometimes named this shape. One expert noted "this looks like a conga line" while another called it a "snake". In these cases, participants defaulted to using mostly verbal communication.

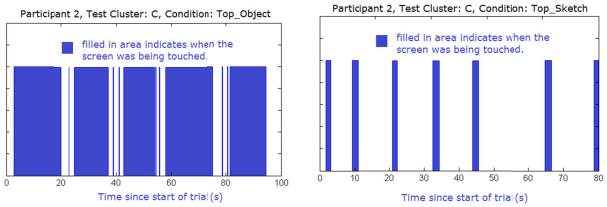

Figure 13. Typical examples of screen contact timings for two similar tasks (Left) using object based interaction (Right) using sketch based interaction.

Strategies for Object based Manipulation

Typically object manipulation only saw the use of two strategies. Either the expert would position one box and ask the worker to move it, or position all four boxes and then have the worker place them all. The one-at-a-time approach seemed to have benefits and weaknesses. On one hand, it allowed for boxes to be placed and used as a reference point for the placement of the other boxes; however, often this would also become a source of occlusion depending on how the worker moved around the scene. In a few cases, the expert would be distracted by the worker and then ask them to wait until the cues were all positioned. This meant that most pairs of participants ended up on the all-at-once and then adjust strategy. One particular pair had a different strategy to the others though. Instead of being mostly silent in the object mode, this pair was very verbally collaborative; the expert used the

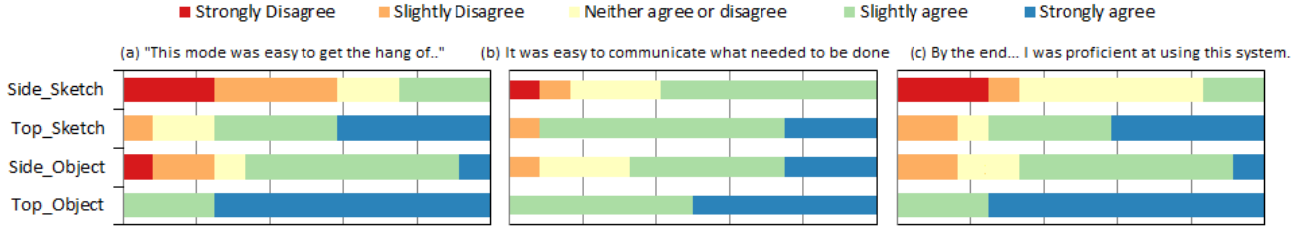

Figure 14. Participant Responses to Qualitative Survey.

worker to guide the position of the cue. A typical example was the following:

> Expert: "Okay, now this needs to be closer to the gray line"
> Worker: "Follow my finger and I'll lead you to the gray line".

Sketching Strategies

The sketching condition saw a multitude of strategies for drawing and specifying the boxes. In this case, one-at-a-time was more common than all-at-once, but the method in which these were specified took the form of either one, two or four lines per box. Figure 15 shows the most common strategies in order of popularity, while the one line method (informally) appeared to be the most efficient. Each of these strategies also seemed to have benefits and limitations.

In the 'Four Lines' approach, the expert could not easily replicate the dimensions of the box- especially from the angled view point, with one participant exclaiming "let me draw that again, that's not supposed to be a diamond, I can't draw rectangles properly". In the 'One Line' approach the worker would align the center of the box with the line. However, the length of the line was often longer than the box, thus verbal adjustments along the line needed to be made. The two lines approach took two forms; either they would draw a 'T' as pictured in Figure 15, or they would draw an 'L' to specify a corner. These were effective methods, though the lines were often not perpendicular, and the correct meaning of the sketch needed confirmation in each case. Other notable strategies that lead to long trial completion times included drawing pairs of vertical and horizontal lines whose intersection points indicated the corners. Several pairs used the markings in the environment to align the positions with the boxes e.g. "between these two gray lines".

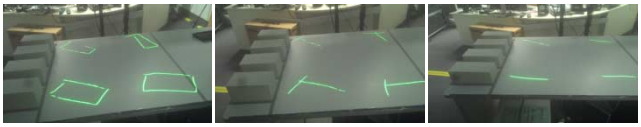

Figure 15. Example Sketching Strategies (left) Four Lines, (center) Two Lines, and (right) One Line

DISCUSSION

A summary of the quantitative user study results can be seen in Table 2. It is apparent that the object manipulation technique was able to improve upon sketching, but not in all circumstances.

Condition	Measure	Object Only	Sketch Only
Top View	Total Time	=	
	Translation	✓	
	Rotation	=	
	Aggregate	✓	
Side View	Total Time		✓
	Translation	=	
	Rotation		✓
	Aggregate	=	

Condition	Measure	Top View	Side View
Object Only	Total Time	✓	
	Translation	✓	
	Rotation	✓	
	Aggregate	✓	
Sketch Only	Total Time	=	
	Translation	=	
	Rotation	=	
	Aggregate	=	

Table 2. Summary of Quantitative Results (Left) Object vs. Sketch, keeping View constant (Right) Top vs. Side Camera View, for each Interaction Technique. '✓' indicates the condition that performed significantly better, while '=' indicates no significant difference.

When using the top camera view, object manipulation provided a significant benefit to translation accuracy whilst taking the same amount of time as sketching. However, in the side angled camera, we see that the object manipulation method was significantly less effective in terms of time and rotation. The right hand side of Table 2 highlights that the change of view point has an effect on the performance of the object manipulation technique whereas sketching was not shown to vary in the same way. Many of these quantitative results differed from the participant self evaluations which highlighted a distinct preference for the object manipulation technique.

Effect of Viewpoint

We originally expected sketching to be hindered by the angled view point, and thus would perform significantly worse than sketching from the top camera. This was based on the assumption of a higher cognitive load making it difficult for the expert to compensate for the perspective change. The difference in difficulty was reported in the questionnaire with many participants stating **side_sketch** was hard to get the hang of and that it was difficult to communicate what needed to be done. In the experimenter observations, it was also noted that many participants verbally exclaimed during the trial 'oh this is hard' or 'hold on, I'm trying to figure out how to adjust'. However, in the quantitative results, despite the perceived difficulty, there was no significant difference in total time spent nor accuracy between **side_sketch** and **top_sketch**. The difference between quantitative and qualitative results can perhaps be explained by the 'learnability' of the system; despite initially having a hard time processing the perspective change, participants also reported that, by the end of the trial, they felt more proficient at sketching from the side.

In the self assessment of the participants, and unlike **side_sketch**, **side_object** was reported as being easy to get the hang of and easy to communicate. Thus, qualitatively speaking, there was no significant disadvantage in using the side view camera.

The quantitative results showed that **side_object** performed worse than **top_object** on all four measures. Firstly, the task time was significantly higher from the side, and this was informally observed to be because the expert would spend a lot of time making minor rotational adjustments to the cue. However, despite all the rotational adjustments, **side_object** was still significantly less accurate than **top_object**. The inference from these results is that object manipulation does not overcome the issues arising from a perspective change.

Effect of Interaction Technique

The experimenter noted that often when object manipulation took more time than sketching it may have been due to the expert spending more time adjusting for minor rotational differences using the system. Rather, while in the sketching mode, most adjustment requests occurred verbally. The measured data agrees with this, and we can surmise that the time spent adjusting rotations was not particularly efficient.

Qualitatively, most participants preferred object manipulation over sketching, which indicates that despite **side_object** performing differently to **side_sketch**, the expert did not perceive this difference.

Factors of High Cost

There were several factors that were not quantified but may have had an effect on the cost measurements within these experiments. One of these is the latency inherent in any networked collaborative system. The crux of the issue is there always exists some noticeable delay between the user touching the screen and appropriate feedback appearing in the form of a projection through the video feed. In our system, this issue is noticeable by the expert but not the worker. The majority of experts in our user study explicitly stated that this latency affected the usability of the

system. However, the effect of this latency was disproportionately noticeable between conditions. Specifically, the majority of participants reported that the latency impaired the use of object manipulation while only some reported an effect on the sketching mode. This may be explained by previous familiarity and confidence with sketching.

Another factor is the effect of occlusion. From the side viewpoint, the positions of the boxes would often occlude the visibility (for the expert) of the cues being projected on the table, but this was not much of an issue in the top view. Additionally, the worker's body and movement would also be a source of occlusion that affected the ability of the expert in both views. As a strategy to deal with this, in the object conditions, many pairs resorted to positioning all four box targets before the worker moved any of them. This strategy essentially made the remote guidance very sequential and could almost be described as a form of asynchronous collaboration. In that form of collaboration, the worker and the expert do not need to work together simultaneously, and instead could collaborate with a significant time separation. This could be useful for certain industrial applications; consider for example, the scenario where a manager connects to the remote site at the start of the day, assigns orders (such as highlighting the places to drill) and moves onto other work without needing to supervise the entire time. Later, the manager could reconnect to assign new orders or make adjustments.

CONCLUSIONS AND FUTURE WORK

We have extended the concept of object-proxy based manipulation for spatially augmented remote guidance. We have implemented a prototype system that allows an expert to use multi-touch gestures to translate, rotate and annotate an object on a video feed causing meaningful projections into the real world.

A user study showed that object manipulation is able to improve upon sketching based SAR remote guidance from a view point perpendicular to the work surface. The study also found that the object based method was less efficient in rotational tasks from an angled viewpoint.

We now suggest some practical improvements to the prototype system, and finish by identifying a new range of research questions to be explored.

The prototype we designed presents a method of SAR remote guidance for spatial arrangement tasks. The primary limitation of this system was the effect of latency on the usability of the system. In any practical implementation, the latency of communication will have a significant impact on the expert. One way of dealing with this could be to provide some immediate local feedback to the expert, realizing that this could potentially cause synchronization issues between the verbal and projected guidance. The idea of dual feedback in collaborative environments has been shown to be effective by Gutwin et al. [10] and insights of that research could be brought into the SAR domain.

Future research may explore the range of tasks that object based methods are suitable for. In the user study we conducted, the tasks were limited to the spatial arrangement of boxes constrained to a plane. There is future work to be undertaken that looks at spatial arrangement in three dimensions, for example, allowing for boxes to be stacked on each other. As a starting point, work conducted by Adcock et al. [2] looks at SAR visual cues for 3D positions of viewpoints (which is, in itself, a challenge for SAR). These cues could be mapped to objects instead. Alternatively, object manipulation based SAR remote guidance could be applied to procedural and psychomotor tasks of the kind described by Henderson and Feiner [11].

Finally, the potential for future work exists in bringing more real-world constraints into the virtual scene. Greater semantic knowledge of objects within the scene could be used to impose further real-world constraints on the virtual targets. The appropriate constraints to use for various task contexts is now an open question for exploration and research.

ACKNOWLEDGMENTS

We sincerely thank: the user study participants, Warren Muller for his expert statistics advice, Henry Gardner and Chris Gunn for their support, and the SUI'14 anonymous reviewers for their helpful comments and feedback.

REFERENCES

1. Adcock, M., Anderson, S., and Thomas, B. RemoteFusion: real time depth camera fusion for remote collaboration on physical tasks. *Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry - VRCAI '13*, ACM Press (2013), 235–242.
2. Adcock, M., Feng, D., and Thomas, B. Visualization of off-surface 3D viewpoint locations in spatial augmented reality. *Proceedings of the 1st symposium on Spatial user interaction - SUI '13*, ACM Press (2013), 1.
3. Bimber, O. and Raskar, R. *Spatial augmented reality: Merging real and virtual worlds*. AK Peters Ltd, 2005.
4. Chen, T., Zhu, Z., Shamir, A., Hu, S.-M., and Cohen-Or, D. 3-Sweep. *ACM Transactions on Graphics 32*, 6 (2013), 1–10.
5. Dragicevic, P., Ramos, G., Bibliowitcz, J., Nowrouzezahrai, D., Balakrishnan, R., and Singh, K. Video browsing by direct manipulation. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, ACM Press (2008), 237.
6. Fussell, S., Setlock, L., Yang, J., Ou, J., Mauer, E., and Kramer, A. Gestures Over Video Streams to Support Remote Collaboration on Physical Tasks. *Human-Computer Interaction 19*, 3 (2004), 273–309.
7. Gauglitz, S., Lee, C., Turk, M., and Höllerer, T. Integrating the physical environment into mobile remote collaboration. *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services - MobileHCI '12*, ACM Press (2012), 241.
8. Gunn, C. and Adcock, M. Using Sticky Light Technology for Projected Guidance. *OzCHI*, (2011), 131–134.
9. Gurevich, P., Lanir, J., Cohen, B., and Stone, R. TeleAdvisor: a versatile augmented reality tool for remote assistance. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, ACM Press (2012), 619.
10. Gutwin, C., Benford, S., Dyck, J., Fraser, M., Vaghi, I., and Greenhalgh, C. Revealing delay in collaborative environments. *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, ACM Press (2004), 503–510.
11. Henderson, S.J. and Feiner, S.K. Augmented reality in the psychomotor phase of a procedural task. *2011 10th IEEE*

International Symposium on Mixed and Augmented Reality, IEEE (2011), 191–200.

12. Hiura, S., Tojo, K., and Inokuchi, S. 3-D tele-direction interface using video projector. *Proceedings of the SIGGRAPH 2003 conference on Sketches & applications in conjunction with the 30th annual conference on Computer graphics and interactive techniques - SIGGRAPH '03*, ACM Press (2003), 1.
13. Kholgade, N., Simon, T., Efros, A., and Sheikh, Y. 3D object manipulation in a single photograph using stock 3D models. *ACM Transactions on Graphics 33*, 4 (2014), 1–12.
14. Kirk, D. and Stanton Fraser, D. Comparing remote gesture technologies for supporting collaborative physical tasks. *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, (2006), 1191.
15. Kondo, D., Kurosaki, K., Iizuka, H., Ando, H., and Maeda, T. View sharing system for motion transmission. *Proceedings of the 2nd Augmented Human International Conference on - AH '11*, ACM Press (2011), 1–4.
16. Kuzuoka, H. Spatial workspace collaboration: A sharedview video support system for remote collaboration capability. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*, ACM Press (1992), 533–540.
17. Lanir, J., Stone, R., Cohen, B., and Gurevich, P. Ownership and control of point of view in remote assistance. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, ACM Press (2013), 2243.
18. Liu, J., Au, O.K.-C., Fu, H., and Tai, C.-L. Two-Finger Gestures for 6DOF Manipulation of 3D Objects. *Computer Graphics Forum 31*, 7 (2012), 2047–2055.
19. Marner, M.R., Irlitti, A., and Thomas, B.H. Improving procedural task performance with Augmented Reality annotations. *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, IEEE (2013), 39–48.
20. Martinet, A., Casiez, G., and Grisoni, L. The effect of DOF separation in 3D manipulation tasks with multi-touch displays. *Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology - VRST '10*, ACM Press (2010), 111.
21. Martinet, A., Casiez, G., and Grisoni, L. Integrality and separability of multitouch interaction techniques in 3D manipulation tasks. *IEEE transactions on visualization and computer graphics 18*, 3 (2012), 369–80.
22. Nacenta, M.A., Baudisch, P., Benko, H., and Wilson, A. Separability of spatial manipulations in multi-touch interfaces. *Proceedings of Graphics Interface 2009*, Canadian Information Processing Society (2009), 175–182.
23. Ou, J., Chen, X., Fussell, S.R., and Yang, J. DOVE : Drawing over Video Environment. .
24. Potter, R.L., Weldon, L.J., and Shneiderman, B. Improving the accuracy of touch screens: an experimental evaluation of three strategies. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88*, ACM Press (1988), 27–32.
25. Ranatunga, D., Feng, D., Adcock, M., and Thomas, B. Towards object based manipulation in remote guidance. *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, IEEE (2013), 1–6.
26. Reisman, J.L., Davidson, P.L., and Han, J.Y. A screen-space formulation for 2D and 3D direct manipulation. *Proceedings of the 22nd annual ACM symposium on User interface software and technology - UIST '09*, ACM Press (2009), 69.
27. Rekimoto, J. SmartSkin: An Infrastructure for Freehand Manipulation on Interactive Surfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02*, ACM Press (2002), 113.
28. Rosenthal, S., Kane, S.K., Wobbrock, J.O., and Avrahami, D. Augmenting on-screen instructions with micro-projected guides: When it Works, and When it Fails. *Proceedings of the 12th ACM international conference on Ubiquitous computing - Ubicomp '10*, ACM Press (2010), 203.
29. Sodhi, R., Benko, H., and Wilson, A. LightGuide: projected visualizations for hand movement guidance. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, ACM Press (2012), 179.
30. Sodhi, R.S., Jones, B.R., Forsyth, D., Bailey, B.P., and Maciocci, G. BeThere: 3D mobile collaboration with spatial input. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, ACM Press (2013), 179.
31. Suenaga, T., Umeda, T., and Kuroda, T. A Tele-Instruction System for Ultrasound Tele-diagnosis. *ICAT ' 99*, (1999), 84–91.
32. Tecchia, F., Alem, L., and Huang, W. 3D helping hands: a gesture based MR system for remote collaboration. *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry - VRCAI '12*, ACM Press (2012), 323.
33. Tsimeris, J.A. Visual Cues for the Instructed Arrangement of Physical Objects Using Spatial Augmented Reality (SAR). 2010, 1–145. https://www.cis.unisa.edu.au/wiki/Tsimeris-minorthesis.
34. Wellner, P. and Freeman, S. *The Double Digital Desk: Shared Editing of Paper Documents (Technical Report EPC-93-108)*. Cambridge UK, 1993.
35. Zheng, Y., Chen, X., Cheng, M.-M., Zhou, K., Hu, S.-M., and Mitra, N.J. Interactive images: cuboid proxies for smart image manipulation. *ACM Transactions on Graphics 31*, 4 (2012), 1–11.