How About the Mentor? Effective Workspace Visualization in AR Telementoring

Chengyuan Lin*[1], Edgar Rojas-Muñoz[1], Maria Eugenia Cabrera[1], Natalia Sanchez-Tamayo[1], Daniel Andersen[1], Voicu Popescu[1], Juan Antonio Barragan Noguera[1], Ben Zarzaur[2], Pat Murphy[2], Kathryn Anderson[2], Thomas Douglas[3], Clare Griffis[3], and Juan Wachs[1]

[1]Purdue University, [2]Indiana University School of Medicine, [3]Naval Medical Center Portsmouth

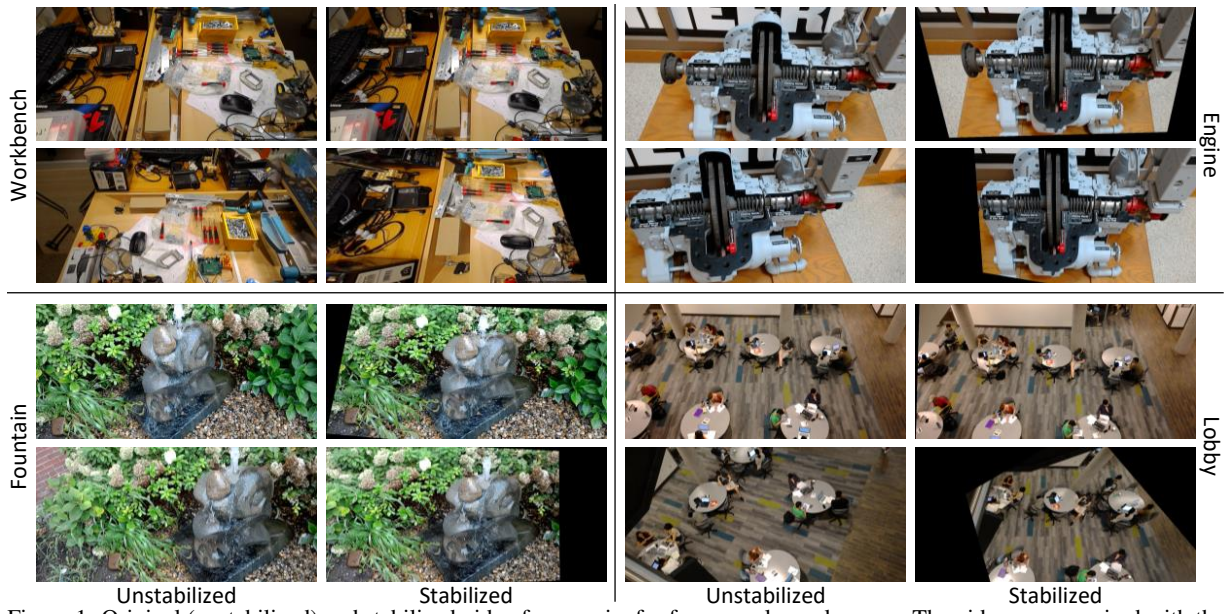


Figure 1: Original (unstabilized) and stabilized video frame pairs for four sample workspaces. The videos are acquired with the camera built in an AR HMD worn by a user who walks around and rotates their head. Our method alleviates the view changes in the original first-person videos, which results in a stable visualization of the workspace, suitable for a remote collaborator, e.g. a mentor. Our method can handle complex 3D geometry (all examples), large view changes (*Workbench, Lobby*), large depths (*Lobby*), and dynamic geometry, complex reflectance properties, and outdoor scenes (running *Fountain*).

Abstract

Augmented Reality (AR) benefits telementoring by enhancing the communication between the mentee and the remote mentor with mentor authored graphical annotations that are directly integrated into the mentee's view of the workspace. An important problem is conveying the workspace to the mentor effectively, such that they can provide adequate guidance. AR headsets now incorporate a front-facing video camera, which can be used to acquire the workspace. However, simply providing to the mentor this video acquired from the mentee's first-person view is inadequate. As the mentee moves their head, the mentor's visualization of the workspace changes frequently, unexpectedly, and substantially. This paper presents a method for robust high-level stabilization of a mentee first-person video to provide effective workspace visualization to a remote mentor. The visualization is stable, complete, up to date, continuous, distortion free, and rendered from the mentee's typical viewpoint, as needed to best inform the mentor of the current state of the workspace. In one study, the stabilized visualization had significant advantages over unstabilized visualization, in the context of three number matching tasks. In a second study, stabilization showed good results, in the context of surgical telementoring, specifically for cricothyroidotomy training in austere settings.

Index Terms: Human-centered computing—Visualization; Graphics systems and interfaces—Mixed / augmented reality.

1 Introduction

As science and technology specialize ever more deeply, it is more and more challenging to gather in one place the many experts needed to perform a complex task. Telecollaboration can transmit expertise over large geographic distances promptly and effectively [28].

A special case of telecollaboration is telementoring, where a mentee performs a task under the guidance of a remote mentor. One approach is to rely only on an audio channel for the communication between mentor and mentee. Telestrators add a visual channel—the mentor annotates a video feed of the workspace, which is then shown to the mentee on a nearby display [30]. The challenge is that the mentee has to switch focus repeatedly away from the workspace, and to remap the instructions from the nearby display to the actual workspace, which can lead to a high cognitive load on the mentee, and ultimately to task completion delays and even errors [4]. Augmented Reality (AR) technology can solve this problem by directly integrating the annotations into the mentee's field of view. The mentee sees the annotations as if the mentor actually drew them on the 3D geometry of the workspace, eliminating focus shifts [2].

*lin553@purdue.edu

A problem less studied but nonetheless of great significance is conveying the workspace to the remote mentor effectively [7, 8]. One approach is to acquire the workspace with an auxiliary video camera, and to send its video feed to the mentor [19]. The approach requires additional hardware, and the auxiliary camera captures the workspace from a different view than that of the mentee. Effective telementoring requires the mentor to see what the mentee sees for the instructions to be as relevant and easy to understand as possible [14]. For example, the mentor might annotate a part of the workspace that is not visible to the mentee due to occlusions, or, conversely, the mentor might not see the part the mentee is working on.

With the advancement of AR, self-contained optical see-through head mounted displays (HMDs) are now available. Such HMDs typically incorporate a front-facing camera, which can capture the workspace from a viewpoint close to the mentee's viewpoint. However, simply providing the mentee first-person video to the mentor is insufficient for effective telementoring [11]. As the mentee changes head position and view direction, the mentor's visualization of the workspace changes frequently and substantially, which adversely affects the mentor's understanding of the scene. This in turn degrades the quality of the guidance provided by the mentor, and ultimately the mentee's performance. For example, when the mentee looks to the left, the workspace visualization shifts by hundreds of pixels to the right; when the mentee moves to the other side of the workspace as might be needed for best access during task performance, the visualization rolls 180°, which results in an upside-down visualization that is frustratingly difficult to parse. What is needed is a robust stabilization of the mentee first-person video, such that it can provide an effective visualization of the workspace to the mentor. The needed *high-level* stabilization has to neutralize the effects of substantial rotations and translations of the acquisition camera, and cannot be provided by prior work *low-level* stabilization techniques that remove jitter in hand-held acquired video.

In this paper we present the design, implementation, and evaluation of a method for robust high-level stabilization of a video feed acquired from a mentee's first-person view, in order to provide a remote mentor with an effective visualization of the mentee's workspace. The output visualization has to be (1) stable, i.e. to show the static parts of the scene at a constant image location, (2) real-time, i.e. to keep up with the input feed, and (3) of high quality, i.e. without distortions, tears and other artifacts. In addition to conveying the workspace to the mentor, the output visualization should also be a (4) suitable canvas on which the mentor can author annotations to provide guidance. The paper investigates three approaches and adopts projective video texture-mapping onto a planar proxy of the workspace geometry, as the approach that best satisfies the design requirements. Fig. 1 illustrates the robustness of our stabilization method on a variety of challenging workspaces.

We evaluated the effectiveness of our stabilization method in two controlled within-subject user studies. One study (n = 30) investigated workspace visualization quality by asking participants to find matching numbers in a video of a workspace annotated with numbers. The study used three workspaces: a *Sandbox*, a *Workbench*, and an *Engine* (the *Workbench* and the *Engine* are shown in Fig. 1 without the numbers). In the control condition, participants watched the original (unstabilized) video acquired with the HMD camera; in the experimental condition, the video was stabilized with our method, which showed significant advantages in terms of task performance and participant workload. For the sandbox workspace we compared our method to a perfectly stable video acquired from a tripod, and there were no significant differences in performance. The second study tested our method in the context of surgical telementoring, where participants (n = 20) practiced cricothyroidotomy (cric) procedures on patient simulators (Fig. 2). The study was conducted in an austere setting of an empty room, with the patient simulator on the floor, with poor visibility achieved with a fog machine, and with loud combat-like noises. Compared to audio-based telementoring, the stabilized video telementoring improved surgical performance significantly. We also refer the reader to the accompanying video.

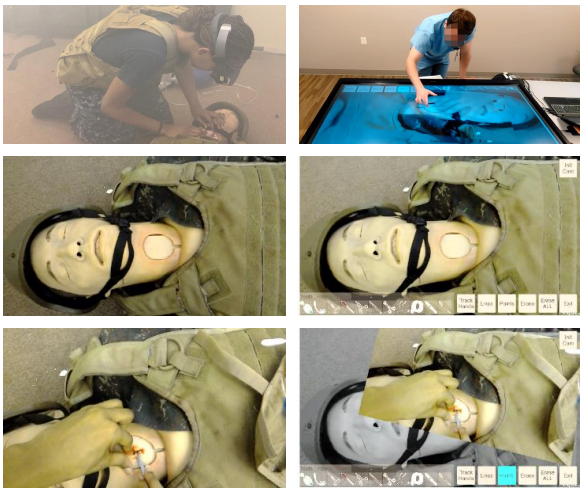


Figure 2: Cricothyroidotomy training in austere environment using video feed stabilized with our method. The mentee wears an AR HMD that acquires the surgical field (top left), the video feed is sent to the mentor where it is stabilized (rows 2-3, raw left, stabilized right), the mentor annotates the stabilized feed (top right), and the annotations are sent to the mentee where they are displayed with the AR HMD. The first frame (grayscale) is used for context.

2 Prior Work

The widespread availability of digital cameras and of broadband internet connectivity enable telecollaboration by acquiring the local workspace with a video camera whose feed is transmitted to a remote site. An important design decision is where to place the camera in order to provide an effective remote visualization of the workspace.

One approach is to mount the camera on a tripod. This approach was used to build a surgical telementoring system where the operating field was acquired with a ceiling-mounted overhead camera [19]. The top view is substantially different from the mentee's view, which reduces telementoring effectiveness, as a mentor can best guide a mentee when the mentor sees what the mentee sees, and when the mentor issues instructions in the mentee's frame of reference. Another surgical telementoring system acquires the operating field with the front-facing camera of a computer tablet mounted with a bracket between the mentee and the patient [3]. The operating field is acquired from a view similar to that of the mentee, but the tablet creates workspace encumbrance. A shortcoming common to both systems is that the operating field is acquired from a fixed view. A second approach is to rely on the local site collaborator to acquire the workspace with a hand-held video camera, changing camera pose continually for a good visualization for the remote collaborator [13]. The problem is that the local collaborator becomes a cameraman, which hinders collaboration.

A third approach is to rely on a head mounted camera [22]. This brings freedom to the local collaborator, who can focus more on the task. A 360° video camera captures more of the environment and provides the remote collaborator with more awareness of the local space [20]. One disadvantage is having to wear the head mounted camera. The disadvantage has been alleviated as internet-connected cameras have been miniaturized, e.g. telecollaboration using Google Glass [27]. We have adopted this third approach. In our context, having to wear the head-mounted camera is not an additional concern since the mentee already has to wear an AR HMD.

The fundamental challenge of acquiring the workspace with a

head-mounted camera is that the visualization of the workspace provided to the mentor changes abruptly, substantially, and frequently as the local collaborator moves their head during task performance. Such a visualization can lead to a loss of situational awareness, to a high cognitive load, to task performance delays and errors, and to cybersickness. Researchers have investigated addressing this challenge by attempting to stabilize the video such that it does not change as the local collaborator moves their head.

One approach of stabilization is to use optical flow to track features over the sequence of frames, to define homographies between consecutive frames using the tracked features, to register all frames in a common coordinate system, and to stabilize each frame by 2D morphing it to the common coordinate system [22]. A second approach is to acquire a 3D geometric model of the workspace, to track the video camera, and to projectively texture map the model with the video frames, from a constant view. One option for acquiring the model is SLAM [13], another option is to use real-time active depth sensing. As we designed our stabilization technique, we investigated both of these approaches, as discussed in Sect. 3.2.

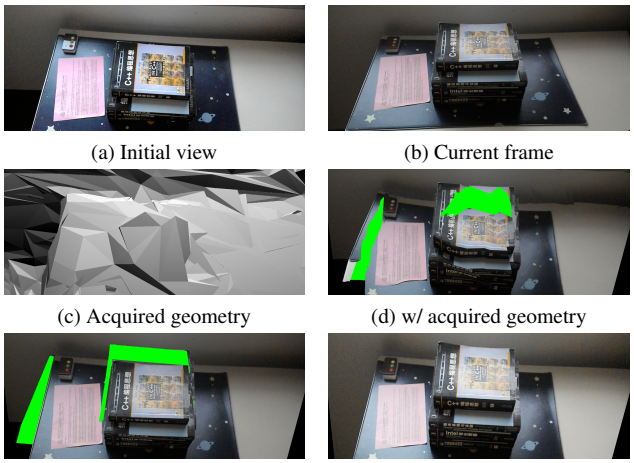
Researchers have developed low-level video stabilization techniques designed to remove small, high-frequency camera pose changes, such as the jitter of a hand-held camera [24, 31], or of a bicycle helmet mounted camera [18]. However, the large amplitude camera pose changes remain. If a hand-held camera is rolled 30°, low-level stabilization preserves the 30° roll, striving for a smooth angle change from 0° to 30°. In contrast, high-level stabilization aims to remove the 30° roll altogether.

Beyond technical challenges, researchers have also investigated video telecollaboration design from a user perspective, to optimize collaboration effectiveness. The problem of obtaining a good view of the workspace has been studied extensively in the context of telemedical consultation [32], where fixed, head-mounted, or hand-held cameras, 2D (view dependent) or 3D (view independent) interfaces each have advantages and disadvantages. A recent study finds that giving remote collaborators independent views is more beneficial than letting the local participant choose the view for the remote participant [17]. The benefit of view independence were also noted in the context of shared live panorama viewing [21], and remote instruction of cockpit operation [12]. Another study found that a scene camera was preferred in video telecollaboration over a head-mounted camera, not just by the remote helper who enjoyed the stable, comprehensive view of the workspace, but also by the worker who preferred not having to wear the camera [11].

Researchers have also demonstrated the acquisition of a complex environment with simple hardware, such as a tablet and its camera [13], to allow a remote collaborator a view-independent exploration of the environment; however, such systems are limited to static environments. Some systems allow the remote collaborator to suggest placement of objects in the workspace [33], again, under the assumption of an otherwise static environment. Complex dynamic scenes are handled by doing away with geometry acquisition, under the assumption that the entire scene is sufficiently far away, which enables panorama acquisition and rendering [26], but this precludes nearby workspaces. Finally, dynamic geometry has been handled through the volumetric fusion of data acquired with multiple off-the-shelf depth cameras, which affords a remote collaborator an independent visualization of the workspace [1]; however, this comes at the cost of additional hardware, intractable in austere environments, and it is limited to the outside-looking-in scenario.

3 High-Level Stabilization of First-Person Video

Consider the AR telementoring scenario with a mentee wearing an optical see through AR HMD. The HMD has a built-in front-facing video camera that captures what the mentee sees. The goal is to use this video feed to inform a remote mentor of the current state of the workspace. In addition to audio instructions, the mentor also provides guidance through graphical annotations of the workspace.



(a) Initial view (b) Current frame

(c) Acquired geometry (d) w/ acquired geometry

(e) w/ true (manual) geometry (f) w/ planar proxy

Figure 3: Stabilization of current frame (b) to initial view (a) by projective texture-mapping onto acquired (c, d), truth (e), or proxy geometry (f). Disocclusion errors are highlighted in green.

Therefore, the video feed should also serve as a canvas on which the mentor authors annotations of the workspace.

3.1 Effective Mentor-Side Visualization Requirements

An effective mentor-side workspace visualization has to satisfy the following requirements:

Stability. The visualization of the workspace should not move, to allow the mentor to examine it in detail. Complex tasks require for the mentor to concentrate on the workspace, and unexpected changes in the visualization are particularly frustrating, forcing the mentor to abandon the AR-enabled graphical communication channel, and to take refuge in the trusted audio communication.

View agreement. The mentor's view of the workspace should be similar to that of the mentee, for the mentor to provide guidance directly in the mentee's context, avoiding any remapping that could confuse the mentee. Furthermore, different viewpoints could show different parts of the workspace to the mentor and mentee, which impedes communication when one party refers to workspace elements not visible to the other party.

Real time. The visualization of the workspace should be up to date, as latency leads to workspace inconsistencies between mentor and mentee, complicating communication.

High visual quality. The visualization should be free of static and temporal artifacts such as tears, holes, and distortions. Of particular importance are scene lines, which should project to lines in the visualization. This is essential for the mentor's ability to understand and annotate the workspace.

3.2 Approaches Considered

Acquiring the workspace with a fixed camera satisfies the stability requirement, but not the view agreement one. A mentee-acquired first-person video satisfies the view agreement requirement, and it is well suited for austere environments since it does not require additional equipment. However, meeting the stability requirement is challenging. As the mentee looks away from the workspace, e.g. to grab a tool, the mentor's visualization changes abruptly and significantly. We investigated three stabilization approaches.

The first is a 2D stabilization approach similar to the one described by Lee and Höllerer [22] , based on tracking and stabilizing 2D video features. The approach lacked robustness in our context, with occasional incorrect feature tracking causing unacceptable stabilization artifacts. The second approach is based on the acquisition of workspace geometry (Fig. 3). Real-time acquisition of

complex 3D scenes is imperfect, resulting in stabilized frame distortions (Fig. 3d); furthermore, the workspace has to be acquired from multiple viewpoints to avoid disocclusion errors (Fig. 3e).

3.3 Stabilization by Projection on Planar Proxy

The third approach investigated, which we adopted, is to projectively texture map the tracked video feed onto a planar approximation of the workspace geometry. The planar proxy is defined once per session. Rendering the textured planar proxy takes negligible time, even on the thinnest of mentor platforms, such as a computer tablet or a smartphone, so the visualization is real time. The visualization is of high quality (Fig. 3f), i.e. without distortions due to inadequate geometric approximation, and without tears due to disocclusion errors. All scene lines project to lines in the visualization. The effect is similar to a photograph of a photograph of a 3D scene. The concatenation of an additional projection does not make the visualization confusing, the same way a visualization makes sense to two or more users seeing it on a display, with no one assuming the true viewpoint from where it was rendered.

4 Theoretical Visualization Stability Analysis

The two possible sources of visualization instability are workspace geometry approximation error, and video camera tracking error. In this section we provide a theoretical analysis of the impact of these two errors on visualization stability. In the next section we provide empirical measurements of visualization stability.

4.1 Visualization instability definition

Given a 3D workspace point P, an initial frame F_0 with view V_0, and a current frame F_i with view V_i, we define the reprojection error of P as the distance $e_i(P)$ between where P should be seen from V_0 and where it is actually seen in the stabilized F_i. In Equation 1, the actual location of P in the stabilized frame is denoted with $\chi(P,V_i,V_0)$, and the correct location $\pi(P,V_0)$ is obtained by projecting P with V_0. The approximate projection function χ depends on the stabilization approximation errors. $e_i(P)$ is relative to the frame's diagonal d to obtain an adimensional, image resolution independent measure of reprojection error.

$$e_i(P) = \frac{\|\chi(P,V_i,V_0) - \pi(P,V_0)\|}{d} \quad (1)$$

Given a point P and two consecutive frames F_i and F_{i+1}, we define visualization instability at P as the absolute change in reprojection error from F_i to F_{i+1}, as given by Equation 2.

$$\varepsilon_i(P) = |e_{i+1}(P) - e_i(P)| \quad (2)$$

4.2 Simulation scenario

We analyze visualization instability in a typical telementoring scenario. The workspace is 1m×1m wide, and it is 1m above the floor (Fig. 4a). This is the largest workspace size for which the mentee can work in the outside looking in scenario—for larger workspaces the mentee would have to travel from one area to another, and stabilizing the mentor view to a single view is not applicable. The actual workspace geometry is in between two planes (dotted lines) that are 20cm apart. This height variation is sufficient to model a workbench with tools on it. The workspace geometry is approximated with the solid line rectangle. The mentee is 1.8m tall, and their default view, to which the video is stabilized, is shown with the black frustum.

We consider two typical mentee view sequences. The first sequence is a 25° pan to the left (blue frustum in Fig. 4a), as needed, for example, to reach for a tool placed just outside the workspace. The panning sequence also has a small lateral translation of 10cm, to account for the translation of the eyes when someone turns their head to the side. The second sequence corresponds to the mentee moving to the corner of the workspace to see it diagonally (green frustum in Fig. 4a), which implies a 50cm lateral translation from the initial position, while looking at the center of the workspace.

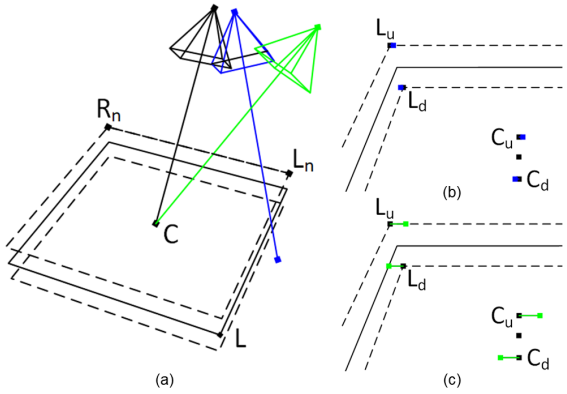


Figure 4: Visualization stability analysis through simulation.

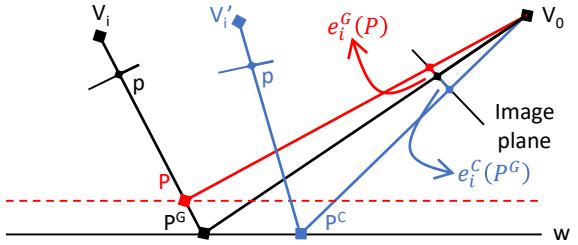


Figure 5: Reprojection error $e_i^G(P)$ due to workspace geometry approximation error, and $e_i^C(P^G)$ due to camera tracking error.

Instability depends on frame to frame view changes. We assume the sequence is completed in 1s, which implies 30 frames at 30Hz. This is a conservative upper bound for the view change speed. For abrupt focal point changes, the mentee does not want to and cannot focus on the workspace during the transition, so any instability will not be perceived, as also noted in walking redirection research that takes advantage of saccadic eye movement to manipulate the visualization [6].

4.3 Dependence on Geometry Approximation Error

In Fig. 5, point P is acquired by video frame V_i and projected onto the proxy plane w at P^G. P and P^G project at different locations onto the stabilized view V_0, which results in the reprojection error $e_i^G(P)$. The dependence of visualization instability on geometry approximation error is obtained by plugging into Equation 2 the expression for χ given in Equation 3, where $V_iP \cap w$ is P^G in Fig. 5.

$$\chi(P,V_i,V_0) = \pi(V_iP \cap w, V_0) \quad (3)$$

The instability induced by geometry approximation error is largest where the true location of a workspace point is farthest from the proxy plane, i.e. on the dotted rectangles in Fig. 4. Fig. 4 illustrates the reprojection errors at the center C and corner L of the workspace proxy, for the last frames of the panning (Fig. 4b) and translation (Fig. 4c) sequences. The correct projections of L_u, L_d, C_u, and C_d are shown with black dots. The actual projections are shown with blue dots for the panning sequence, and with green dots for the translation sequence. As expected, the reprojection error is tiny for the panning sequence since the viewpoint translation is minimal. Pure panning would have a zero reprojection error. The maximum instability at the center of the workspace (i.e. C in Fig. 4) is 0.03% and 0.17% for the panning and translation sequences, respectively. The maximum is reached for the last frame of the sequence, where the viewpoint translation is largest. For an HDTV display with a diagonal of 2,200 pixels and 1m in length, the instability figures translate to 1.1pix and 0.5mm for the panning sequence, and 5.5pix and 2.5mm for the translation sequence. We computed the maximum instability over the entire workspace to be 0.05% and 0.25% for the two sequences, respectively, which occurs at the workspace corners,

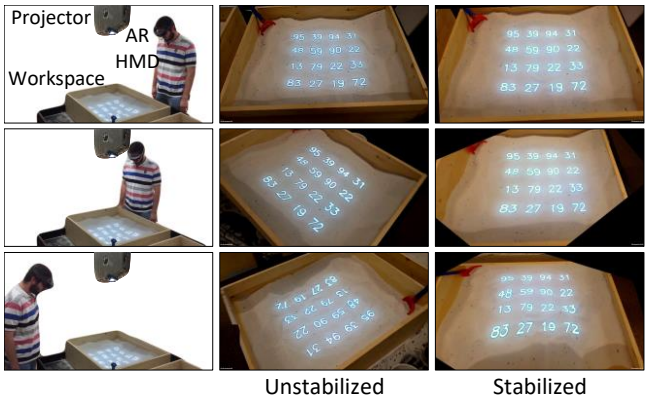


Figure 6: Sandbox workspace with overhead projected numbers acquired with video-camera built into an AR HMD (left column), original, unstabilized video frame (middle), and stabilized video frame (right).

i.e. L_n and R_n in Fig. 4a, for the last frame.

An important advantage of our method is that the geometric approximation is constant, i.e. the proxy plane does not change. This means that, when the mentee translates their viewpoint, the instability is not only small, but also smooth, and when the mentee pauses to focus on a part of the workspace, the instability is 0. For a method that uses a geometric model acquired in real time, the instability is noisy, even when the mentee does not move.

4.4 Dependence on Camera Tracking Error

The second source of visualization instability is the error in tracking the video camera which acquires the workspace. Using Fig. 5 again, let us now assume that proxy plane point P^G is an actual workspace point to factor out all geometry approximation error. P^G is captured at pixel p by the frame with true viewpoint V_i. If V_i is incorrectly tracked at V_i', then p is incorrectly projected onto the proxy at point P^C, which generates reprojection error $e_i^C(P^G)$. The dependence of visualization instability on camera tracking error is obtained by plugging into Equation 2 the expression for χ given by Equation 4, where w is the workspace proxy.

$$\chi(P,V_i,V_0) = \pi(V_i' p \cap w, V_0) \tag{4}$$

Unlike for the instability due to the workspace geometry approximation, tracking inaccuracy affects the entire frame uniformly. We have measured tracking accuracy to be 2 degrees for rotations and 2cm for translations. In the scenario above, these maximum tracking errors translate to a 2.68% and a 1.45% instability, figures that dwarf the instability caused by geometric approximation error (Sect. 4.3). Even assuming tracking that is an order of magnitude more accurate than what our AR HMD provides, instability due to geometry approximation will still be smaller than instability due to tracking.

In conclusion, we have defined instability metrics to be used in the empirical validation, and we have established that instability due to geometric error is dwarfed by that due to camera tracking error, which validates, at principle level, our approach.

5 User Study I: Number Matching

We developed a method for stabilizing the video of a workspace captured by a head mounted camera. The stabilized video serves as a visualization of the workspace for a remote collaborator. In a first controlled user study, we tested the effectiveness of workspace visualization by asking participants to find matching numbers in the original and the stabilized videos, for three workspaces.

5.1 Experimental Design

Participants. We recruited participants ($n = 30$, 8 female) from the graduate student population of our university, in the 24–30 age group. We opted for a within-subject design, with each participant performing the task in all conditions.

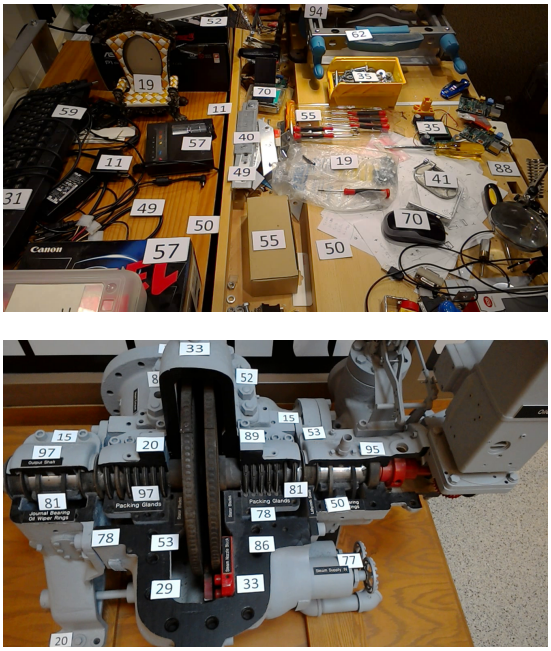


Figure 7: *Workbench* (top) and *Engine* workspaces used in study I.

Task. A participant was seated 2m away from an LCD monitor with a 165cm diagonal. The monitor displays a video of a workspace annotated with numbers, and the participant is asked to find pairs of matching numbers. When a participant spots a matching pair, they call out the number, and an experimenter tallies the number of matches found. All numbers called out by participants were correct matches, i.e. they were not just reading out numbers at random.

Workspace 1: Sandbox. The first workspace is a sandbox in our lab (Fig. 6). The sandbox is approximately 1m×1m in size, and it is placed about 1m off the floor. The sand had a depth variation of about 20cm, so this corresponds to the scenario investigated by the theoretical instability analysis in Sect. 4. An overhead projector displays a matrix of 4×4 numbers on the sandbox. The workspace was acquired with the front-facing camera of an AR HMD (i.e. Microsoft's HoloLens [25]) worn by an experimenter who walked around the sandbox while looking at its center. The experimenter starts out at the default position, where the numbers are correctly oriented (first row of Fig. 6). This is also the view to which the video was stabilized. The experimenter occasionally pans the view to the side. Then the experimenter walks to the corner of the sandbox (second row of Fig. 6), and even on the other side, which makes the numbers appear upside down in the video (third row of Fig. 6). This results in a video sequence where the matrix of numbers moves considerably. The video shows 21 matrices, and each matrix was shown for 5s, for a total video length of 105s. 18 of the 21 matrices had exactly one pair of matching numbers, and 3 of the matrices had no matching numbers. Half the numbers of two consecutive matrices are the same, which means that when the video switches from one matrix to the next, exactly 8 of the 16 numbers change. All 8 numbers change simultaneously at the end of the 5s. When a matrix had a matching pair, at least one of the numbers in the pair was replaced for the next matrix, such that a matching pair would not persist longer than the 5s that each matrix is displayed.

Workspace 2: Workbench. The second workspace is an actual workbench cluttered with tools (Fig. 1 and Fig. 7). The acquisition path was similar to that for the *Sandbox* workspace. The tallest tool reached 30cm above the workbench plane. The experimenter wearing the AR HMD impersonating a mentee started out at the default position, then panned the view, and then finally moved to

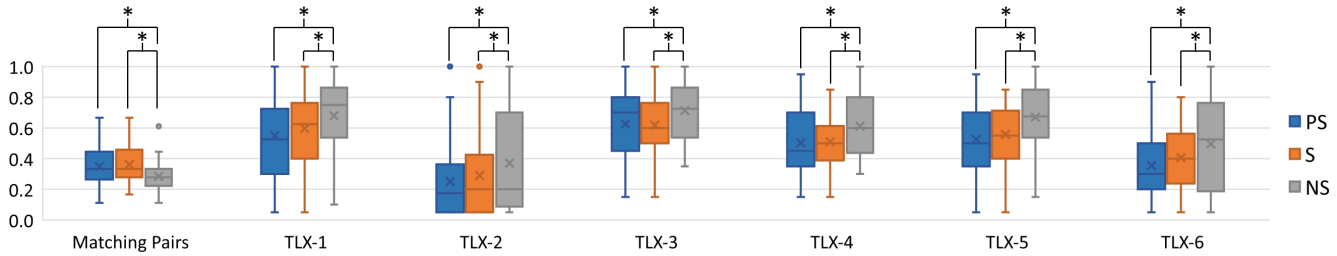


Figure 8: Normalized box and whisker plot of pairs found, and of NASA-TLX subscores, for each of the three *Sandbox* conditions: perfectly stabilized (PS), stabilized (S), and not stabilized (NS). The star indicates significance ($p \leq 0.05$). No S to PS difference was significant.

the side of the workbench to see it from a direction rotated by 90°. The numbers were added to the workspace using pieces of paper, all facing the mentee in the initial position. There were 24 numbers, 8 of which appeared twice, so 8 numbers were unique. Although the numbers on paper did not change, the mentee moved tools on the workbench covering and uncovering a few numbers. Furthermore, as the mentee viewpoint translated, some of the numbers would appear and disappear due to occlusions.

Workspace 3: Engine. The third workspace is an Engine mounted on the floor, 80cm high (Fig. 1 and Fig. 7). The *Engine* was decorated with numbers and was acquired similarly to the *Workbench*.

Conditions. Each participant performed the number matching task for the *Sandbox* workspace in each of three conditions, in randomized order. In one control condition, the participant was shown the raw video with no stabilization (NS). In a second control condition, the participant was shown a perfectly stable (PS) video that was acquired by placing the AR HMD on a mannequin head mounted on a tripod at the default position. In the experimental condition, the participant was shown the video stabilized with our method (S). The hypotheses related to the *Sandbox* were that (1) participants will perform better in the S condition compared to the NS condition, and that (2) participants will not perform better in the PS condition compared to the S condition. A subgroup of 20 participants were tested for each of the *Workbench* and the *Engine* workspaces, for each of two conditions. Participants were shown the original, unstabilized video in the control condition, and the stabilized video in the experimental condition.

Metrics. We measured participant task performance as the number of pairs found. We also measured participant workload using the NASA Task Load Index (NASA-TLX) questionnaire [15], and participant simulator sickness using the Simulator Sickness Questionnaire (SSQ) [16]. Better performance means more matching pairs found, lower cognitive load, and absence of simulator sickness.

5.2 Results and Discussion

A within-subject statistical analysis compared the three *Sandbox* conditions, with three data points for each metric and for each participant. The participants and the order of the trials were treated as blocks in the statistical design. The data normality assumption was confirmed with the Shapiro-Wilk test [29]. In addition, the data equal-variance assumption was confirmed with the Levene test [23], so no data transformation was needed. We ran a repeated measures ANOVA [9] with Bonferroni correction [5] for each condition pair, i.e. PS vs NS, PS vs S, and S vs NS. The two conditions for the *Workbench* and *Engine* were similarly compared, except that no Bonferroni correction is needed.

Fig. 8 gives the box and whisker plot [10] of the number of pairs found, and of the six NASA-TLX subscales, for each of the three *Sandbox* conditions. The six subscales are: mental demand, physical demand, temporal demand, performance, effort, and frustration. All seven metrics are normalized. The plot indicates the inter-quartile range (IQR) with a box, the average value with an x, the median value with a horizontal line, farthest data points that are not outliers with whiskers, and outliers with dots. Outliers are data points "outside the fences", i.e. more than 1.5 times the IQR from the end of the box. NS participants found on average 28% or 5.1 of the 18 matching pairs. S participants found on average 36% or 6.5. PS participants found on average 34% or 6.3. The differences between S and NS, and PS and NS are significant, while the difference between PS and S is not. The best performing participant found 12 of the 18 matching pairs for both the S and PS conditions, performance levels that are within the fence and therefore not outliers; this participant only found 8 matching pairs in the NS condition.

Table 1: Comparison between the number of pairs found in the no stabilization (NS) and stabilization (S) conditions.

Workspace	NS	S	S - NS	p-value
Workbench	5.45±0.83	5.95±1.19	0.50±0.28	0.043*
Engine	5.05±1.57	6.10±1.29	1.05±0.31	0.002*

Table 2: p-values of NASA TLX subscore differences between no stabilization (NS) and stabilization (S) conditions (i.e. NS-S).

Workspace	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
Workbench	0.000*	0.000*	0.001*	0.188	0.356	0.001*
Engine	0.005*	0.050*	0.000*	0.034	0.002*	0.001*

S and PS participants reported significantly lower cognitive load than those in NS on all six NASA-TLX subscales, and there was no significant difference between PS and S. For NS, the upper fence exceeded the maximum possible value of 1.0, and it was therefore capped at 1.0, for all six NASA-TLX subscales. This indicates the high cognitive load in the NS condition, and it eliminates the possibility of outliers. For S and PS, two of the scales had the upper fence at 1.0, which leaves the possibility of outliers for the other four scales. However, there was only one outlier for each of the PS and S conditions, both for the TLX-2 scale, which increases the confidence that PS and S place less demand on the participant.

Table 1 gives the number of pairs found for the *Workbench* and the *Engine* workspaces, for each of the unstabilized (NS) and the stabilized (S) conditions. S has a significant advantage for both workspaces. Table 2 compares the NASA TLX scores between the S and NS conditions (i.e. NS-S, as lower NASA TLX scores indicate less demand on the participant). Most S advantages are significant. For the *Sandbox* workspace, the analysis of the Total Severity score derived from the SSQ answers indicates the absence of simulator sickness in all three conditions. Furthermore, there are no significant differences for any of the three differences PS-NS, S-NS, and PS-S, for any SSQ subscore. While this suggests that our stabilization might not induce simulator sickness, and that discomfort levels are similar to those for a perfectly stabilized video, the absence of differences between PS and NS indicates that the exposure might have been too short and the workspace too simple for a revealing simulator sickness comparison between the three conditions.

The SSQ provided more insight in the case of the more visually complex *Workbench* and *Engine* workspaces (Table 3). S had a

Table 3: p-values of SSQ Total Severity score differences between no stabilization (NS) and stabilization (S) conditions (i.e. NS-S).

Workspace	Nausea	Oculomotor	Disorientation	Total Severity
Workbench	0.019*	0.001*	0.116	0.004*
Engine	0.053	0.060	0.019*	0.021*

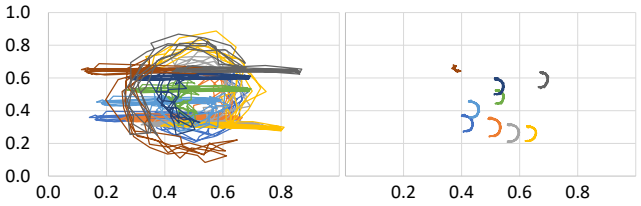


Figure 9: Trajectories of 9 tracked feature points, in normalized pixel coordinates, for the NS (left) and S (right) *Sandbox* conditions.

significant advantage over NS in terms of Total Severity score, for both workspaces. The S advantage was due to less nausea and oculomotor effort for the flatter but more cluttered *Workbench*, and due to disorientation for the more occlusion/disocclusion prone *Engine*. Although the differences between conditions were significant, for no workspace and no condition did the Total Severity score increase from pre- to post- exposure above the threshold of 70, which would indicate the presence of simulator sickness.

5.3 Empirical Visualization Stability Analysis

Sect. 4 defined visualization instability and analyzed its dependence on the workspace geometry approximation error and on the camera tracking error. Here we measure the actual instability in the raw video and in the stabilized video by tracking nine salient feature points over the entire *Sandbox* sequence. The features are dark particles mixed in with the white sand, and they cover the matrix area uniformly. The frame trajectories of the tracked features are shown in Fig. 9, where the coordinates in the 1,280×720 video frame were normalized. Whereas the tracked points move considerably in the NS video, their trajectory is short and smooth in the S video. The average reprojection error (Equation 1) over all feature points and all frames is 13.5%±7.9% for NS and 2.0%±1.8% for S; the maximum reprojection error is 37.5% for NS and 5.8% for S.

The average visualization instability (Equation 2) over all 9 feature points is given in Fig. 10 for both the unstabilized and the stabilized sequences. These instability values are based on empirical values for the $\chi(P,V_i,V_0)$ and $\pi(P,V_0)$ from the definition of reprojection error Equation 1. Instability is large for NS, and it is largest for the first part of the sequence, when the mentee panned their head left and right repeatedly. This is expected since, for a non-stabilized sequence, panning motions change the frame coordinates of workspace features quickly and substantially. Instability is low for our stabilized sequence, and it is lower for the first part of the sequence when workspace geometry approximation error has little influence on instability. For the first part of the sequence, the instability is very low most of the time, with the exception of some small spikes which we attribute to camera tracking latency. The average instability is $0.081\% \pm 0.082\%$ for the NS sequence, and about eight times lower for the S sequence at $0.011\% \pm 0.0093\%$.

6 User Study II: Austere Surgical Telementoring

We conducted a second user study, which tests the benefits of stabilization in the context of a complete surgical telementoring system. The mentee acquires the surgical field with a front-facing video camera built into their AR HMD, the video is transmitted to the remote mentor site, the video is stabilized, the stabilized video is shown to the mentor, the mentor provides guidance by annotating the stabilized video, and the annotations are sent to the mentee site, where

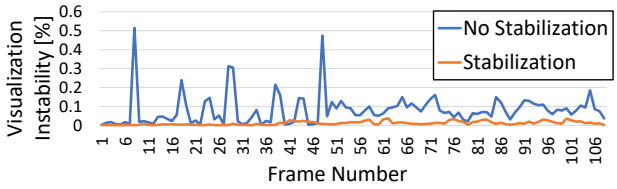


Figure 10: Empirical visualization instability measured by tracking feature points over the video sequences.

they are overlaid onto the surgical field using the AR HMD. The study evaluates the benefit of stabilization indirectly: the hypothesis is that the stabilized video leads to a better mentor understanding of the operating field, to better guidance for the mentee, and ultimately to better mentee performance.

6.1 Experimental Design

Participants. The participants served as mentees in the study. We recruited participants ($n = 20$) from the corpsmen of a naval medical center who were training for performing surgical procedures in austere settings. The participant age range was 18–43, and 3 participants were female. The study used two mentors that are teaching faculty at a surgery residency program. The mentor site was 900km away from the mentee site. We opted for a within-subject design, with each participant performing a task in both conditions.

Task. The participants performed a practice cric on a synthetic patient simulator in an austere setting (Fig. 2). The cric is an emergency procedure performed when a patient is not able to breathe due to airway obstruction. The procedure entails performing precise incisions through multiple layers of neck tissue, opening up the cricoid cartilage, inserting and securing a breathing tube, and connecting a breathing bag to the tube. Since emergent, the procedure stands to benefit greatly from telementoring.

Conditions. In the experimental condition (EC), the mentee benefited from visual and verbal guidance from the mentor. The visual guidance was provided through the AR HMD, which overlaid mentor-authored annotations onto the operating field, such as freehand sketched incision lines, or dragged-and-dropped instrument icons. The mentor monitored the operating field and authored annotations based on a first-person video of the operating field acquired by the mentee, which was stabilized with our method. In the control condition (CC), the mentee benefited from verbal mentor guidance.

Metrics. The mentee performance was evaluated by two expert surgeons located at the mentee site. The experts used the cric evaluation sheet typically used at the naval center to score the performance of the mentees. The evaluation sheet contains 10 subscales based on procedure steps, which are scored with a 5-level Likert Scale. The subscales evaluate aspects related to anatomical landmark identification, incision performance, and patient airway acquisition. The overall mentee performance score was computed as the average of the 10 subscale scores.

6.2 Results and Discussion

A within-subject statistical analysis was run to compare both conditions, with two data points for each metric and for each participant. The condition was treated as an independent variable, while each of the expert evaluation scores were treated as dependent variables. The participants and the order of the trials were treated as blocks in the statistical design. The data normality and equal variance assumptions were confirmed with the Shapiro-Wilk [29] and the Levene test [23], respectively, and a repeated measures ANOVA was run [9].

The results are shown in Fig. 11, which gives means and standard deviations. The total performance score (EE-T) was significantly higher ($p = 0.04$) for EC than for CC. The means for each of the ten subscale scores (i.e. EE-1 to EE-10) favor EC over CC, but only two of the differences are significant, i.e. for EE-8 ($p = 0.03$) and for

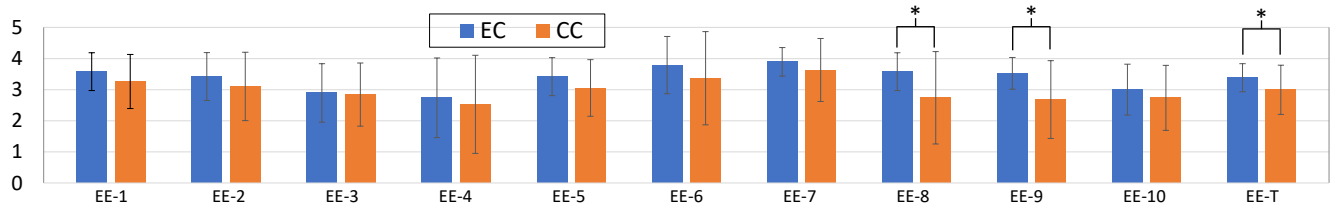


Figure 11: Procedure subscale (EE-1 to EE-10) and overall (EE-T) cric performance. EC has an advantage over CC for each metric. The star indicates a significant advantage ($p \leq 0.05$).

EE-9 ($p = 0.01$). We attribute the lack of significance for the score differences for the other subscales to the low number of participants. EE-8 verifies that the cuff of the Melker canula was inflated with 10ml of air, which indicates that there is air circulating through the tube. EE-9 verifies that the air actually makes it into the lungs of the patient (simulator) as indicated by a bilateral rise and fall of the chest. On the other hand, EE-10 verifies that the cannula is properly secured with tape for patient transport, so it concerns a step beyond the end of the actual cric, and participants could score highly on EE-10 even if the procedure actually failed. Thus, EE-8 and EE-9 are important scores that depend on the success of all previous steps, and they validate the entire procedure.

The mentee moves their head considerably as they reach for surgical instruments, which causes numerous, substantial, and abrupt changes in the input video. In one typical instance, a mentee translated their head for a total of 7.66m over a 3min and 12s sequence, with spikes of over 20cm per second. In the same sequence, the mentee rotates the view direction by over 1,500°, which is more than four full rotations. These large view changes make the raw video unusable at the mentor, and our stabilization is essential to the success of the AR telementoring system.

The workspace in the surgical telementoring study is highly dynamic, with the mentee's hands and instruments moving in the video feed. While such dynamic environments are challenging for approaches that rely on real time geometry acquisition, the dynamic workspace does not pose any additional challenge to our approach. Note that our definition of instability (Equation 2) does apply to dynamic environments since it does not simply measure how far the projection of a 3D point moves from one frame to the next, which would penalize the moving elements of the environment even in a perfectly stabilized visualization; instead, our definition is based on how far away the 3D point is in the visualization from where it should be in a perfectly stabilized visualization.

7 Conclusion, Limitations, and Future Work

We have presented the design and evaluation of a method for stabilizing a first-person video of a workspace, such that it can effectively convey the workspace to a remote collaborator. We investigated three approaches and we chose an approach that projectively texture maps the registered video feed onto a planar proxy of the workspace. The approach has the advantages of stability, view agreement, real time performance, lack of distortions, lack of disocclusion errors, good temporal continuity, and robustness with workspace geometric, reflectance property, and motion complexity. We refer the reader to the accompanying video for additional stabilization examples.

The stabilized video doesn't always contain all the pixels in the input mentee video. This happens when the mentee view frustum is not a subset of the mentor view frustum. For example, in Fig. 1, for the *Engine* workspace, the top left unstabilized frame captures more of the text on the wall than its stabilized counterpart. This is due to the fact that the mentor view frustum was chosen to encompass tightly the workspace, i.e. the engine. A wider mentor field of view would have kept the entire back wall pixels in the stabilized frame. Certainly, this would come at the cost of a lower resolution on the workspace, and each application should decide what works best in its own context. Another possibility to be explored as future work, is to not insist on a fixed mentor view, but rather a view that slowly keeps up with the mentee view in order to show the mentor everything the mentee sees. For example, if the the mentee chooses to focus on a completely different area of the workspace, the mentor view should gradually focus on that area as well.

When the mentee looks away from the workspace, the mentor's live visualization of the workspace is truncated, or even interrupted if the mentee view frustum is completely disjoint from the mentor view frustum. One solution for mitigating this problem is to rely on previous frame pixels to maintain workspace visualization continuity. Of course, these are not live pixels so they can only be used for orientation purposes, and not for up to date situational awareness. We took this approach in the cric study, where the a previous frame is used to provide context (see frame in Fig. 2, row 3, right). The background frame is shown in grayscale to make it clear to the mentor that it is not a live shot. Future work could explore updating the background frame to keep up with a dynamic workspace, i.e. to be more recent and less obsolete. Another direction of future work is to rely on a series of background images and to rely on an approach similar to projective texture mapping to choose the most suitable background image for the current frame. Suitability can be quantified as the number of missing mentor frame pixels that are filled in, which requires view direction similarity, and as the continuity of the transition from live to background pixels, which requires viewpoint similarity.

One limitation to address in future work is that our first study does not provide a sufficiently long exposure to measure simulator sickness. Another direction of future work is to examine conveying the workspace to the remote collaborator through a Virtual Reality (VR) HMD, where simulator sickness is likely to be a bigger factor.

The second user study compared AR telementoring based on our stabilization to a control condition where the mentor and mentee could only communicate through audio. One reason for this is that audio communication is the most frequently used means of communication between mentor and mentee. The second reason is that the unstabilized video was judged by the expert surgeon mentors as unusable in the context of the emergent cric and of the austere conditions. In other words, it was not possible to run a user study where one of the conditions was AR telementoring with the raw, unstabilized video. Future studies could attempt to isolate the stability factor in settings where the surgical intervention and the environment are less stressful to make the unstabilized video acceptable, at least for the purpose of a user study.

Our work tests AR surgical telementoring with actual health care practitioners, in a real training exercise, in a highly demanding austere setting, towards placing AR technology into societal service.

Acknowledgments

We thank our Augmented Reality Tea group for insightful comments and suggestions. This work was supported by the United States Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-14-1-0042, and by the United States National Science Foundation under Grant DGE-1333468. The views expressed in article, reflect the results of research conducted by the author(s) and do not necessarily reflect the official policy or position of the funders, including but not limited to the Department of the Navy, Department of Defense, or the United States Government.

REFERENCES

[1] M. Adcock, S. Anderson, and B. Thomas. Remotefusion: real time depth camera fusion for remote collaboration on physical tasks. In *Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, pp. 235–242. ACM, 2013.

[2] D. Andersen, V. Popescu, M. E. Cabrera, A. Shanghavi, G. Gómez, S. Marley, B. Mullis, and J. P. Wachs. Avoiding focus shifts in surgical telementoring using an augmented reality transparent display. In *MMVR*, vol. 22, pp. 9–14, 2016.

[3] D. Andersen, V. Popescu, M. E. Cabrera, A. Shanghavi, G. Gomez, S. Marley, B. Mullis, and J. P. Wachs. Medical telementoring using an augmented reality transparent display. *Surgery*, 159(6):1646–1653, 2016.

[4] E. M. Bogen, K. M. Augestad, H. R. Patel, and R.-O. Lindsetmo. Telementoring in education of laparoscopic surgeons: An emerging technology. *World journal of gastrointestinal endoscopy*, 6(5):148, 2014.

[5] C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[6] B. Bridgeman, D. Hendry, and L. Stark. Failure to detect displacement of the visual world during saccadic eye movements. *Vision research*, 15(6):719–722, 1975.

[7] A. Budrionis, G. Hartvigsen, and J. G. Bellika. Camera movement during telementoring and laparoscopic surgery: Challenges and innovative solutions. In *SHI 2015, Proceedings from The 13th Scandinavien Conference on Health Informatics, June 15-17, 2015, Tromsø, Norway*, number 115, pp. 1–5. Linköping University Electronic Press, 2015.

[8] L. Chen, T. W. Day, W. Tang, and N. W. John. Recent developments and future challenges in medical mixed reality. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 123–135. IEEE, 2017.

[9] R. A. Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pp. 66–70. Springer, 1992.

[10] M. Frigge, D. C. Hoaglin, and B. Iglewicz. Some implementations of the boxplot. *The American Statistician*, 43(1):50–54, 1989.

[11] S. R. Fussell, L. D. Setlock, and R. E. Kraut. Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 513–520. ACM, 2003.

[12] S. Gauglitz, C. Lee, M. Turk, and T. Höllerer. Integrating the physical environment into mobile remote collaboration. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pp. 241–250. ACM, 2012.

[13] S. Gauglitz, B. Nuernberger, M. Turk, and T. Höllerer. World-stabilized annotations and virtual scene navigation for remote collaboration. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pp. 449–459. ACM, 2014.

[14] W. W. Gaver, A. Sellen, C. Heath, and P. Luff. One is not enough: Multiple views in a media space. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pp. 335–341, 1993.

[15] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[16] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[17] S. Kim, M. Billinghurst, and G. Lee. The effect of collaboration styles and view independence on video-mediated remote collaboration. *Computer Supported Cooperative Work (CSCW)*, 27(3-6):569–607, 2018.

[18] J. Kopf, M. F. Cohen, and R. Szeliski. First-person hyper-lapse videos. *ACM Transactions on Graphics (TOG)*, 33(4):78, 2014.

[19] H. Kuzuoka. Spatial workspace collaboration: a sharedview video support system for remote collaboration capability. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 533–540. ACM, 1992.

[20] G. A. Lee, T. Teo, S. Kim, and M. Billinghurst. Mixed reality collaboration through sharing a live panorama. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, p. 14. ACM, 2017.

[21] G. A. Lee, T. Teo, S. Kim, and M. Billinghurst. A user study on mr remote collaboration using live 360 video. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 153–164. IEEE, 2018.

[22] T. Lee and T. Höllerer. Viewpoint stabilization for live collaborative video augmentations. In *Mixed and Augmented Reality, 2006. ISMAR 2006. IEEE/ACM International Symposium on*, pp. 241–242. IEEE, 2006.

[23] H. Levene. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pp. 279–292, 1961.

[24] S. Liu, L. Yuan, P. Tan, and J. Sun. Bundled camera paths for video stabilization. *ACM Transactions on Graphics (TOG)*, 32(4):78, 2013.

[25] Microsoft hololens. `https://www.microsoft.com/en-us/hololens`. Accessed: 2019-02-14.

[26] J. Müller, T. Langlotz, and H. Regenbrecht. Panovc: Pervasive telepresence using mobile phones. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10. IEEE, 2016.

[27] B. A. Ponce, M. E. Menendez, L. O. Oladeji, C. T. Fryberger, and P. K. Dantuluri. Emerging technology in surgical education: combining real-time augmented reality and wearable computing devices. *Orthopedics*, 37(11):751–757, 2014.

[28] H. Sebajang, P. Trudeau, A. Dougall, S. Hegge, C. McKinley, and M. Anvari. The role of telementoring and telerobotic assistance in the provision of laparoscopic colorectal surgery in rural areas. *Surgical Endoscopy and Other Interventional Techniques*, 20(9):1389–1393, 2006.

[29] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[30] S. Treter, N. Perrier, J. A. Sosa, and S. Roman. Telementoring: a multi-institutional experience with the introduction of a novel surgical approach for adrenalectomy. *Annals of surgical oncology*, 20(8):2754–2758, 2013.

[31] M. Wang, G.-Y. Yang, J.-K. Lin, S.-H. Zhang, A. Shamir, S.-P. Lu, and S.-M. Hu. Deep online video stabilization with multi-grid warping transformation learning. *IEEE Transactions on Image Processing*, 28(5):2283–2292, 2019.

[32] G. Welch, D. H. Sonnenwald, H. Fuchs, B. Cairns, K. Mayer-Patel, R. Yang, H. Towles, A. Ilie, S. Krishnan, H. M. Söderholm, et al. Remote 3d medical consultation. In *Virtual realities*, pp. 139–159. Springer, 2011.

[33] J. Zillner, E. Mendez, and D. Wagner. Augmented reality remote collaboration with dense reconstruction. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 38–39. IEEE, 2018.