# How About the Mentor? Effective Workspace Visualization in AR Telementoring

Category: Research

Paper Type: algorithm/technique

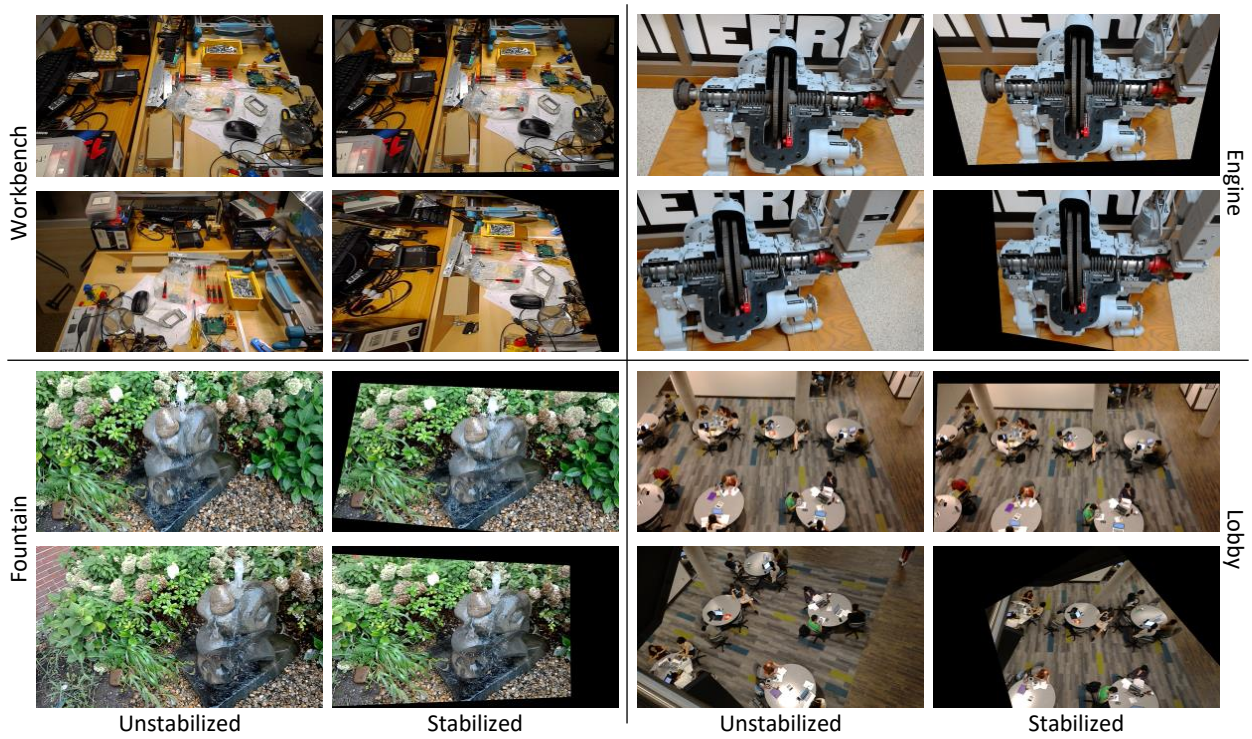



Fig. 1: Original (unstabilized) and stabilized video frame pairs for four sample workspaces. The videos are acquired with the camera built in an AR HMD worn by a user who walks around and rotates their head. Our method alleviates the view changes in the original first-person videos, keeping the static parts of the scene in place, 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) technology 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. A less studied but nonetheless 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 changes head position and view direction, the mentor's visualization of the workspace changes frequently, unexpectedly, and substantially, which hinders mentor performance. This paper presents a method for the robust high-level stabilization of a mentee first-person video to provide an effective 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, for prompt and effective telementoring. The visualization is implemented by projecting the tracked video feed onto a planar proxy of the workspace. The stabilization method was evaluated in two user studies. In one study, stabilized visualization had significant advantages over unstabilized visualization, in the context of three number matching tasks. In a second study, stabilization was tested, with good results, in the context of surgical telementoring, specifically for cricothyroidotomy training in austere settings.

**Index Terms**—Augmented Reality, telementoring, mentor, first-person video, stable visualization

## 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.

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 enhance communication with a visual channel—the mentor annotates a video feed of the workspace, which is then shown to the mentee on a nearby display. 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. 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.

A problem less studied but nonetheless of great significance is conveying the workspace to the remote mentor effectively. One approach is to acquire the workspace with an auxiliary video camera, and to send its video feed to the mentor. Not only does the approach require additional hardware, but the auxiliary camera captures the workspace from a different view than that of the mentee. Effective telementoring requires for the mentor to see what the mentee sees for the instructions to be as relevant and easy to understand as possible. For example, when using an auxiliary camera, 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 technology, self-contained optical see-through head mounted displays (HMDs) are now available. Such HMDs typically incorporate a front-facing onboard 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. 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 *high-level* stabilization of the mentee first-person video, such that it can provide an effective visualization of the workspace to the mentor.

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 reports the investigation of three approaches: (a) 2D stabilization based on tracked frame features, and projective texture-mapping of the tracked video feed on a 3D model of the workspace, with the 3D model being either (b) a 3D triangle mesh or (c) a proxy plane. The proxy plane approach was adopted as the one 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 participants did not perform significantly better in the perfect stable condition. The second study tested our method in the context of surgical telementoring, where participants (n = 20) practiced cricothyroidotomy 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.

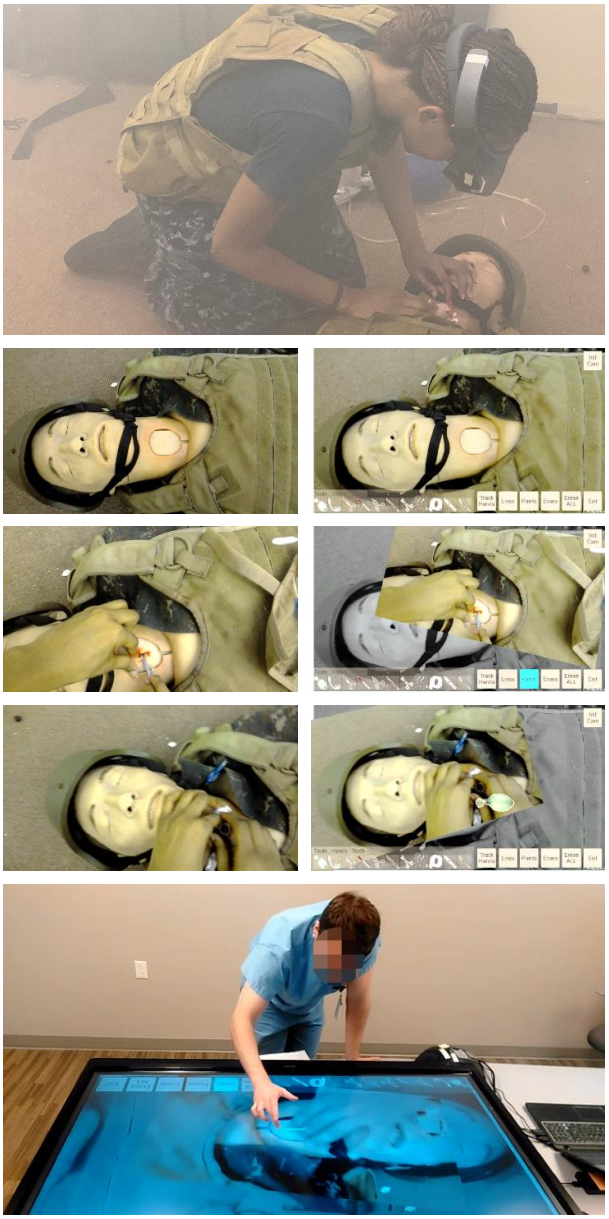


Fig. 2: Cricothyroidotomy training in austere environment using video feed stabilized with our method. The mentee wears an AR HMD that acquires the surgical field (top), the video feed is sent to the mentor where it is stabilized (rows 2-4, raw left, stabilized right), the mentor annotates the stabilized feed (bottom), and the annotations are sent to the mentee where they are displayed with the AR HMD. For this study, the stabilized video was shown over a grayscale background image pre-acquired from the initial view, which provides 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 [2]. 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 [1]. 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 [8]. The problem is that the local collaborator becomes a cameraman, which hinders collaboration.

A third approach is to rely on a head mounted camera [13]. This brings freedom to the local collaborator, who can focus more on the collaborative task. A 360° video camera captures more of the environment and provides the remote collaborator with more awareness of the local space [12]. 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 [20]. 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 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 the coordinate system of the first frame, and to stabilize each frame by 2D morphing it to the constant coordinate system of the first frame [13]. 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 [8], 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.

We point out that low-level video stabilization techniques do not meet our goal of high-level video stabilization. We define low-level video stabilization as the process of removing small, high-frequency camera pose changes, such as the jitter of a hand-held camera [15, 22], or of a bicycle helmet mounted camera [11], with the goal of obtaining a smooth video, as if the video were acquired with a physically stabilized (e.g. gyro-stabilized) camera. The low-level stabilized video keeps the large amplitude pose changes of the acquiring camera. If a hand-held camera is panned 30° to the right, low-level stabilization preserves the 30° pan, striving for a smooth angle change from 0° to 30°. In contrast, high-level stabilization aims to remove the 30° pan altogether.

## 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. 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, due to occlusions, a different viewpoint could show different parts of the workspace to the mentor and mentee, which brings substantial challenges in communication when one party refers to workspace elements not visible to the other party.

*Real time.* The visualization of the workspace should be always up to date. Any latency leads to an inconsistent state of the workspace between the mentor and mentee, which complicates 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. The high visual quality is essential not only for the mentor's understanding of the workspace, but also for the mentor's ability to draw and place graphical annotations accurately.

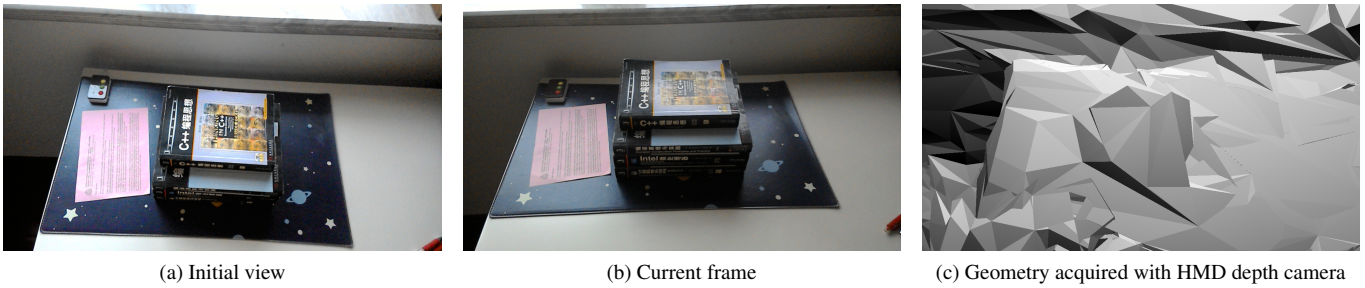
### 3.2 Approaches Considered

The stability concern is well addressed by acquiring the workspace with an additional, fixed camera. One concern is where to mount this additional camera. Meeting the view agreement requirement is difficult, as placing the additional video camera on a tripod such that it captures the workspace from the mentee's viewpoint encumbers the workspace. Placing the camera opposite the mentee to capture the workspace at a similar angle but from the other side is problematic because reprojection to the mentee's viewpoint requires substantial viewpoint translation which results in objectionable disocclusion artifacts. In a previous system we opted for an overhead placement of the additional camera, which does not satisfy the view agreement requirement, and also suffers from occlusions when the mentee leans over the workspace. The need to deploy an additional camera is another shortcoming of this approach, of particular concern in austere settings.

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, the stability requirement disqualifies the mentee first-person video from being used as is. As the mentee temporarily looks away from the workspace, for example to grab a tool, the visualization of the mentor would change abruptly and significantly. Even the small head motions that the mentee makes naturally to focus on various parts of the workspace are sufficiently distracting to decrease mentor performance.

We investigated three approaches for stabilizing the mentee first-person video. The first approach attempts a 2D stabilization in the same spirit of the method described by Lee and Höllerer [13]. Our pipeline finds frame features using SIFT [16], computes an initial feature matching using FLANN [18], removes outlier matches and finds an initial homography using RANSAC [5], optimizes the homography using least squares, and warps the current frame to the first frame using the optimized homography. This 2D stabilization approach did not produce satisfactory results in our contexts. First, the workspace is not truly planar, and the planar approximation computed by the method changes from frame to frame based on the current set of tracked features, which leads to poor stabilization. Second, the method is not sufficiently robust. Occasional incorrect feature tracking leads to large homography errors and poor video stability. Third, when the mentee abruptly looks away from the workspace, e.g. to grab a tool, feature tracking fails altogether, and regaining tracking is difficult. Finally, our application has a very low tolerance for stabilization artifacts—even a single video stabilization artifact can ruin the surgical telementoring experience.

The second approach for mentee first-person video stabilization that we investigated acquires the 3D geometry of the workspace, it tracks the HMD video camera, and it renders from a stable viewpoint the workspace geometry projectively texture mapped with the video frames. In theory, this approach should yield a perfectly stable visualization of the workspace: accurate geometry and an accurately tracked projective video texture create a consistent 3D scene that appears stable when



(a) Initial view (b) Current frame (c) Geometry acquired with HMD depth camera

(d) Frame stabilized using acquired geometry (e) Frame stabilized using accurate geometry (f) Frame stabilized on planar proxy

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

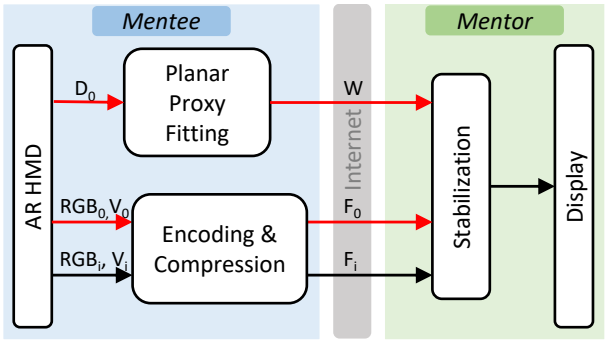



Fig. 4: System overview. Red arrows illustrate the data flow during system initialization, and black arrows during system operation.

rendered from a fixed view. In practice, high-quality real time depth acquisition and geometric modeling of a dynamic scene with complex geometric and reflectance properties remains a challenging problem. Our AR HMD does acquire scene geometry with built-in depth camera, but the quality of the acquired depth is insufficient for a stable image. Any imperfection in the underlying workspace geometric model distorts the visualization. In Fig. 3, the goal is to project the current frame (b) to the initial view targeted by stabilization (a); using the geometry acquired by the AR HMD depth camera (c) results in objectionable distortions (d), which change from frame to frame.

We argue that even if the AR HMD would acquire a flawless RGBD stream, projective texture mapping the acquired depth would still not produce satisfactory results. Whenever a surface is visible from the stabilized viewpoint but not from the acquisition viewpoint, the resulting disocclusion error causes a disturbing tearing artifact in the visualization. Using Fig. 3 again, even projecting onto perfect geometry results in an imperfect visualization suffering from disocclusion errors (e). Disocclusion errors could be addressed by acquiring a complete workspace model by integrating a sequence of RGBD frames acquired with a moving depth camera, e.g. with a Kinect Fusion like approach [19], but such a sequential acquisition is ill-suited for highly dynamic environments with ever-changing geometry. Another option is to acquire the workspace with multiple stationary depth cameras simultaneously, but this complicates the hardware setup substantially, beyond what is practical in an austere setting.

### 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 proxy of the workspace geometry. Fig. 4 gives an overview of our pipeline. The workspace plane $W$ is defined, once per session (red arrows) based on the depth channel $D_0$ of the first frame acquired by the AR HMD. $W$ is sent to the remote mentor and it remains constant for the entire telementoring session. The world coordinate system is defined as the coordinate system $V_0$ of the first frame. The color $RGB_0$ of the first frame is used to define a background image, that will provide context to the visualization of the stabilized frames. $RGB_0$ and $V_0$ are encoded into a frame $F_0$ sent to the mentor during initialization stage.

Then, during operation, the color $RGB_i$ and pose $V_i$ of each frame $F_i$ are transferred to the mentor. The current frame is stabilized by projectively texture mapping the frame over the workspace plane, on top of an optional background image. 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, without tears, and without distortions, with scene lines projecting to lines in the visualization (Fig. 3f). 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 estimates of visualization stability, as measured in our experiments.

### 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 $\chi$ 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.

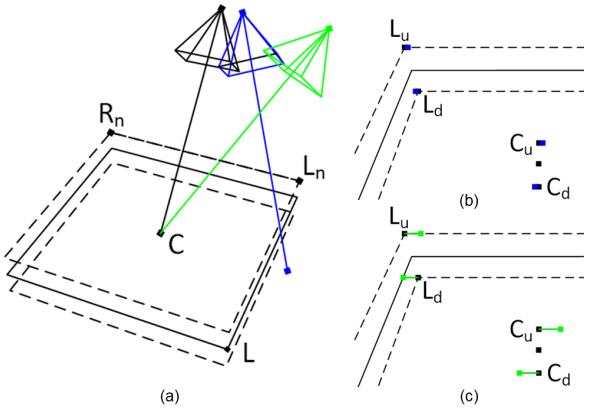



Fig. 5: Visualization stability analysis through simulation.

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

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

In order to estimate the visualization instability of our method, we have simulated a typical telementoring scenario, where the mentee performs a task at a workspace in front of them. In Fig. 5a, the workspace is 1m×1m wide, and it is 1m above the floor. The actual workspace geometry is in between two planes (dotted lines) that are 20cm apart. Our method approximates the workspace geometry with the solid line rectangle. The mentee is assumed to be 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. 5a), 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. 5a). This translation sequence implies a 50cm lateral translation from the initial position, while looking at the center of the workspace.

Instability depends on the amount the view changes from one frame to the next, so the number of frames in the sequence is important. We assume there are 30 frames in the sequence, which, at 30Hz, implies completing the sequence in 1s. This is a conservative upper bound for the view change, assuming the mentee actually examines the workspace during the sequence. When the mentee simply wants to change focal point from the center to the side of the workspace, 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 [4].

### 4.3 Dependence on Geometry Approximation Error

Fig. 6 illustrates the reprojection error caused by geometry approximation error. Workspace 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 $\chi$ given in Equation 3, where $V_iP \cap w$ is $P^G$ in Fig. 6.

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

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. 5. Fig. 5 illustrates the reprojection errors at the center $C$ and corner $L$ of the workspace proxy, for the last frame of the panning sequence (Fig. 5b), and for the last frame of the translation sequence (Fig. 5c). 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. A pure panning sequence would have a zero reprojection error.

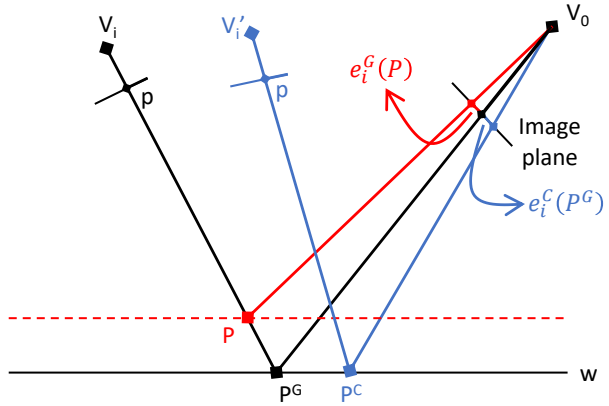



Fig. 6: Reprojection error $e_i^G(P)$ due to workspace geometry approximation error, and $e_i^C(P^G)$ due to camera tracking error.

Table 1 gives the visualization instability for each of the two sequences. The first row gives the maximum instability at the center of the workspace (i.e. $C$ in Fig. 5) over the frames of the sequences. This maximum is reached for the last frame of the sequence, where the viewpoint translation is largest. The second row gives the maximum instability over the entire workspace, which is reached for the last frame and for the near corners of the workspace, i.e. $L_n$ and $R_n$ in Fig. 5a. 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.

An important advantage of our method is that the geometric approximation is constant, i.e. the proxy plane is 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. 6 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 $\chi$ given by Equation 4, where $w$ is the workspace proxy.

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

Unlike for the instability due to the workspace geometry approximation, tracking inaccuracy affects the entire frame uniformly. We

Table 1: Visualization instability due to geometric approximation error for two mentee sequences.

| | Panning | Translation |
|---|---|---|
| Center | 0.03% | 0.17% |
| Max | 0.05% | 0.25% |

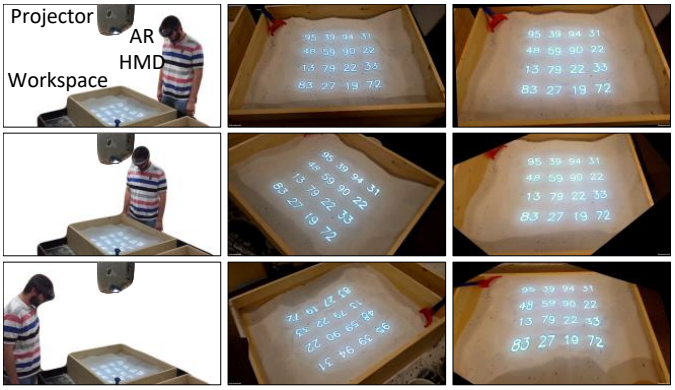



Fig. 7: Sandbox workspace with overhead projected numbers acquired with video-camera built into an AR HMD (left column), original, unstabilized video frame (middle), and stabilized video frame (right).

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 (Table 1). Even assuming tracking that is an order of magnitude more accurate than what our AR HMD provides, the instability due to workspace geometry approximation will still be smaller than the instability due to tracking.

## 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.

*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, so no penalty for incorrect matches was needed; in other words, participants were calling out a number only when they spotted the matching pair, and they were not just reading out the numbers at random in the hope of stumbling upon the correct matching pair by chance.

*Workspace 1: Sandbox.* The first workspace is a sandbox in our lab (Fig. 7). 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 [17]) 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. 7). 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. 7), and even on the other side, which makes the numbers appear upside down in the video (third row of Fig. 7). 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.

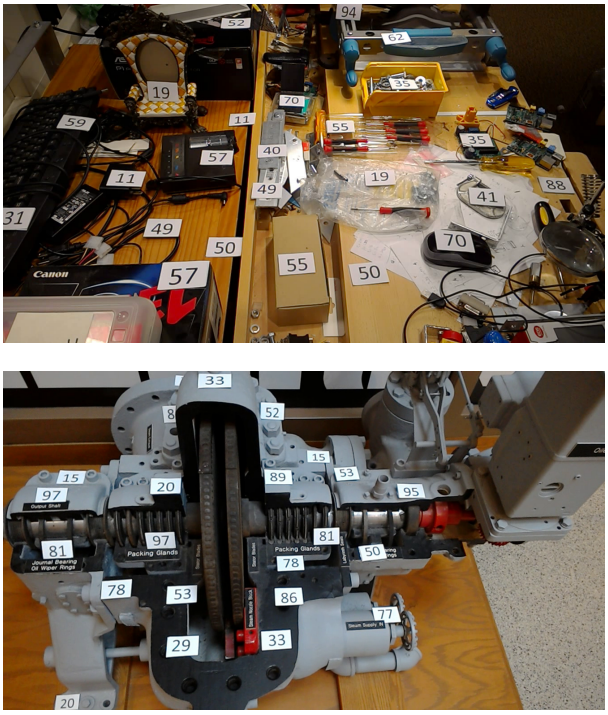


Fig. 8: *Workbench* (top) and *Engine* (bottom) workspaces of study I.

*Workspace 2: Workbench.* The second workspace is an actual workbench cluttered with tools (Fig. 1 and Fig. 8). 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 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 actual Engine mounted on the floor, with a height of 80cm (Fig. 1 and Fig. 8). The *Engine* was decorated with numbers and was acquired similarly to the *Workbench*.

*Conditions.* Each participant performed the number matching task

Fig. 9: Acquisition of perfectly stable video using a tripod.

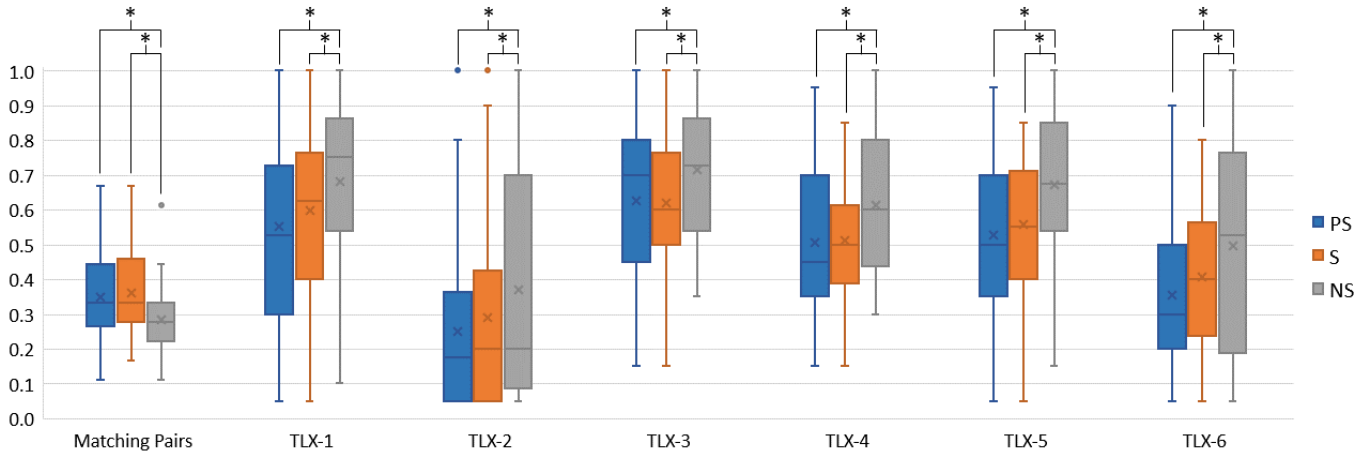



Fig. 10: Normalized box and whisker plot of pairs found, and of NASA-TLX subscores, for each of the three *Sandbox* conditions: perfectly stabilized (PS), stabilized (S), and not stabilized (NS). The star indicates significance ($p \leq 0.05$). No difference between S and PS was significant.

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 (Fig. 9). 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 [9], and participant simulator sickness using the Simulator Sickness Questionnaire (SSQ) [10]. 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 was run to compare the three conditions for the *Sandbox*, with three data points for each metric and for each participant. The condition was treated as an independent variable, while the metrics were treated as dependent variables. 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 [21]. In addition, the data equal variance assumption was confirmed with the Levene test [14], so no data transformation was needed. We ran a one-way ANOVA [6] with Bonferroni correction [3] 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. 10 gives the box and whisker plot [7] 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. As customary, the plot indicates, for each series, 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 NS 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 2: 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 3: 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 NS on all six NASA-TLX subscales, and no subscale reported a 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 the PS condition, and one for the S condition (both for the TLX-2 scale), which strengthens the confidence that PS and S place less demand on the participant.

Table 2 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 3 compares the NASA TLX scores between 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 sick-

Table 4: 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* |

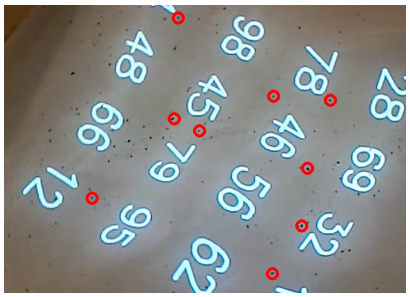


Fig. 11: Features selected for empirical visualization stability analysis, highlighted here with red circles for illustration purposes.

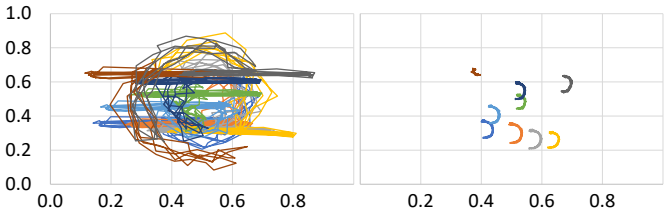



Fig. 12: Trajectories of 9 tracked feature points, in normalized pixel coordinates, for the NS (left) and S (right) *Sandbox* conditions.

ness 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, i.e. Nausea, Oculomotor, and Disorientation. 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 to for a revealing simulator sickness comparison between the three conditions.

The SSQ provided better insight in the case of the more visually complex and realistic *Workbench* and *Engine* workspaces, as shown in Table 4. S had a 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 was the Total Severity score increase from pre- to post- exposure above the threshold of 70, which would indicate the presence of simulator sickness.

Any within-subject study can also be analyzed as a between-subject study with half the data points, by discarding the second half of each participant data. Such a between-subject analysis confirms that the advantages of S over NS are significant, and that there is no significant advantage of PS over S.

### 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 (Fig. 11). 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. 12, 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. 13 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

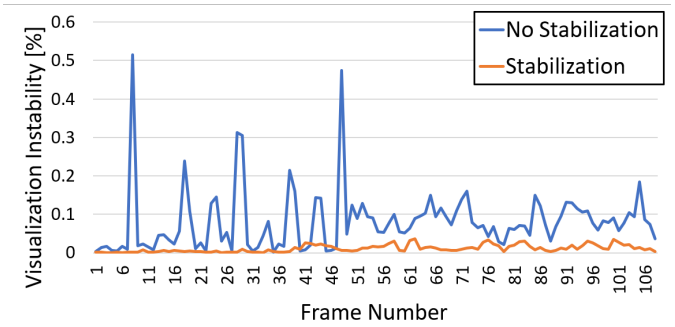



Fig. 13: Empirical visualization instability measured by tracking feature points over the video sequences.

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%±0.082% for the NS sequence, and about eight times lower for the S sequence at 0.011%±0.0093%.

## 6 User Study II: Surgical Telementoring in Austere Settings

The first user study focuses on the benefits of stabilization for workspace understanding, which was tested directly with participants who performed a task based on the workspace visualization. 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 they are overlaid onto the surgical field using the AR HMD. This user study evaluates the benefit of stabilization indirectly, in terms of mentee performance: the hypothesis is that the stabilized video leads to a mentor's better understanding of the operating field, which leads to better guidance for the mentee, which ultimately translates in 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 for each of two conditions.

*Task.* The participants were asked to perform a practice cricothyroidotomy on a synthetic patient simulator in an austere setting (Fig. 2). The cricothyroidotomy is an emergency procedure performed when a patient is not able to breathe through the nose or mouth due to airway obstruction. The procedure entails performing precise incisions through multiple layers of neck tissue, opening up the cricoid cartilage incision, inserting and securing a breathing tube, and connecting a breathing bag to the tube. The procedure stands to benefit greatly from telementoring since often there is no time to bring an expert surgeon to the patient, or to evacuate the patient.

*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

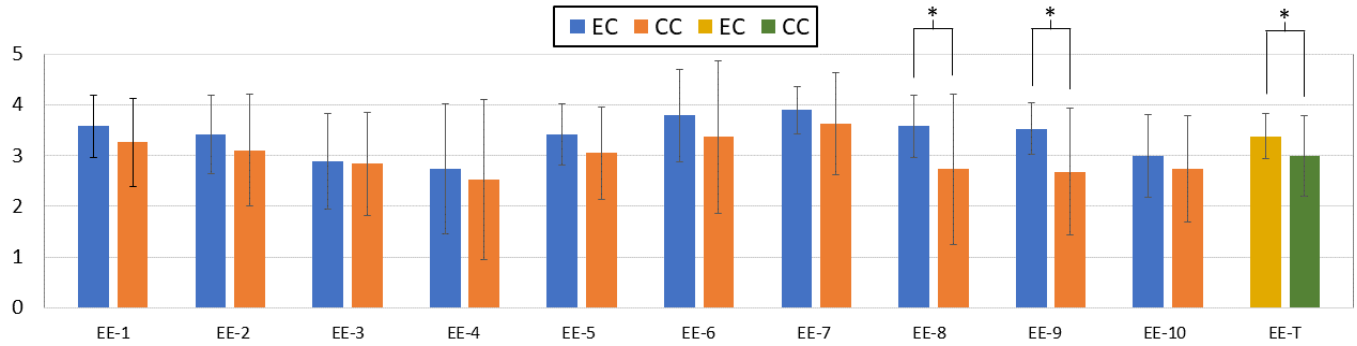



Fig. 14: Procedure subscale (EE-1 to EE-10) and overall (EE-T) cricothyroidotomy performance. EC has an advantage over CC for each metric. The star indicates a significant advantage ($p \leq 0.05$).

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 only from verbal guidance from the mentor.

*Metrics.* The mentee performance was evaluated by two expert surgeons located at the mentee site. The experts used the cricothyroidotomy 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 test [21] and the Levene test [14], respectively, and a one-way ANOVA was run [6].

The results are shown in Fig. 14, 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 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 cricothyroidotomy, 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.

Whereas the workspace used in our number matching study was stationary, in the case of the surgical telementoring study the workspace was highly dynamic, with the mentee's hands and instruments moving in the video feed. While such dynamic environments pose great challenges for alternative approaches that rely on the real time acquisition of workspace geometry, 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 shown in the visualization from where it would be shown 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 video accompanying our paper for additional stabilization examples.

We have formally defined instability and we have performed a theoretical analysis that isolated the dependence of the instability on workspace geometry approximation error, and on camera tracking error. Then we validated our stabilization method in the context of providing an eloquent description of the workspace that is sufficient to perform challenging, cognitively demanding tasks. Finally, we validated our method in a complete AR surgical telementoring system, where it was used to visualize the operating field for a mentor located 900km away.

We conclude that the simplicity, robustness, and visual quality achieved by our stabilization method are hard to match by approaches based on 3D scene acquisition, especially in the context of demanding applications such as surgical telementoring, where instability artifacts cannot be tolerated.

One limitation that will be addressed 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 the scenario where the workspace is conveyed to the remote collaborator through a VR HMD, where simulator sickness is likely to be a bigger factor even for short exposures.

Another limitation of the present work is that 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. Future studies will have to also compare to a control condition where the mentor receives an unstabilized video, over larger number of participants, that might reveal significant differences for more subscales of surgical procedure performance.

Our work focused on AR telementoring, but our stabilization method is likely to be useful in other contexts where a second user needs to see what a first user sees. The current method aims to project the acquired video to a stationary default view. Future work could examine projecting the acquired video to a stable but changing view, e.g. an orbiting camera trajectory circling above the workspace. Finally, our work could be extended to transfer one user's first-person view to the first-person view of a second user, allowing the second user to change the view on the workspace interactively.

Our work tests AR surgical telementoring with actual health care practitioners, in a real training exercise, in a highly demanding austere setting, which takes an important step towards moving AR technology out of labs and placing it at the service of our society.

## REFERENCES


[1] 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.

[2] Anonymous. Withheld for double blind review.

[3] 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.

[4] 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.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

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

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

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

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

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

[18] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009.

[19] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pp. 127–136. IEEE, 2011.

[20] 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.

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

[22] 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.