Look-over-there: Real-World Co-Located Cross-Referencing Using Augmented Reality

Yuqi Zhou and Voicu Popescu

Purdue University, 610 Purdue Mall, West Lafayette, IN, U.S.A.

Keywords: Collaborative Augmented Reality, Transparent Display, Attention Guidance.

Abstract: This paper presents a method that allows a "guide" to point out an element of the real world, i.e., a reference point, to a "tourist". The guide and tourist stand side-by-side and each hold a tablet whose camera is aimed at the scene. The guide annotates the reference point on their tablet, and it is sent and displayed on the tourist’s tablet. Then the device zooms in and guides the tourist towards the target object. A user study shows that the method has significantly shorter reference point localization times compared to a conventional augmented reality interface. Furthermore, the study shows that the method can provide directional guidance through the annotation alone, without any reliance on the visual appearance of the region of the reference point, as needed for challenging scenes with repeated patterns or devoid of visual features.

1 INTRODUCTION

One of the requirements for successful collaboration is that collaborators can point out to each other an element of the 3D scene for common reference. This can be challenging for a variety of reasons. One reason is that natural language can be ambiguous, which makes it difficult to convey precise spatial directions. Another reason is that the working space can be complex and repetitive, which makes it hard to isolate a reference point out of many similar candidates. Augmented reality (AR) allows annotating elements of the real world, which could be used to annotate a reference point in support of collaboration.

An optical see-through AR headset can annotate a reference point directly into the user’s visual field, so they are well suited for indicating the true direction to the reference point. However, AR headsets remain expensive, bulky, dim, and with a limited field of view. Handheld AR displays, implemented by tablets or phones, have the advantages of lower cost, mass deployment, larger field of view, and better robustness with scene lighting conditions. However, the video-see-through AR interface implemented by handheld displays shows the annotation of the reference point on the 2D display, and not directly into the user’s view of the 3D scene. Consequently, the user has to memorize the scene features around the annotation seen on the display, and then, once they shift focus to the scene, find the reference point from memory. In other words, the AR handheld display facilitates, but does not completely eliminate the search for the reference point. To address this issue, one has to transform the frame acquired by the video camera to what the user would see if the tablet were transparent. This user perspective rendering (UPR) of the frame eliminates the need for memorization and improves the accuracy of the AR directional guidance.

In this paper we introduce Look-over-there, a method for enhancing a handheld AR display’s ability to direct the user’s attention to a specific reference point in a real-world setting. Two collaborators stand side-by-side (Fig. 1 b), one (i.e., the guide) pointing out an element of interest to the other (i.e., the tourist). Using handheld devices, such as a tablet or a phone, both collaborators view a live video feed of the scene captured by their device’s camera. The guide annotates the reference point on their display (Fig. 1 a), which appears on the tourist’s display (c). To further improve the accuracy of directional guidance provided to the tourist, the tourist is asked to center the annotation on their display while the visualization zooms in on the scene (c). This refinement stage transforms the display into an UPR transparent display with good alignment between the image on the display and the scene behind it. Once the tourist’s view is correctly aimed at the reference point, the tourist can move the display out of the way to see...
Figure 1: Look-over-there method overview. A guide and a tourist stand side by side (left and right in image b). The guide marks a point of interest on their tablet (red circle in a). The annotation appears on the tablet of the tourist (red circle in c). Although the annotation is at the correct location on the tablet, it is not at the correct location in the visual field of view of the tourist, so the directional guidance provided is inaccurate. During a guidance refinement stage, the tourist tilts up the tablet (d) to center the annotation (e), while the visualization zooms in progressively, from the large field of view of the tablet camera, to the small angle subtended by the tablet in the tourist’s visual field. This improves the accuracy of the directional guidance provided by the tablet, indicating to the tourist the true direction to the reference point, which the tourist then finds with their naked eyes (f). In c and e the background seen by the tourist around the tablet is simulated for illustration purposes.

The reference point directly (f).

We have evaluated our method empirically. The improvement from the refinement stage was confirmed by laboratory experiments: a user study approved by our Institutional Review Board shows that, compared to a conventional AR display, the directional guidance refinement of Look-over-there significantly reduces reference point localization times; furthermore, the study shows that Look-over-there can provide directional guidance through the annotation alone, without any reliance on the visual appearance of the reference point region, which affords directional guidance for difficult scenes that are devoid of visual features or contain repetitive patterns. We have also tested our approach on indoor and outdoor real world scenes, as shown in the accompanying video.

In summary, our paper contributes: (1) a robust and accurate handheld display AR interface for users to indicate a real world scene location to their collaborators; (2) an empirical evaluation of our interface, including in a controlled user study, which confirms the advantages of our interface over conventional AR.

2 PRIOR WORK

The ability to refer to the same element of a shared real-world workspace is essential for successful collaboration. AR technology has the ability to annotate the real world, and thus it can be used to link the real world view of two or more collaborators.

Optical see-through AR interfaces provide guidance in the user’s view of the real world, facilitating the transition for the user to see the target with their naked eyes. When the annotation is visible to the user on the active part of the AR headset, directional guidance is a solved problem. However, optical see-through AR headsets typically have a small field of view, and special interfaces are needed to guide the user’s attention to out-of-sight reference points (Bork et al., 2018). AR headsets remain expensive, bulky, and dim, which currently precludes their use in day-to-day applications. Our work focuses on handheld (video see-through) AR displays.

Handheld video see-through AR displays offer the key benefits of easier entry, as tablets and phones are already ubiquitously deployed. Users can manipulate content on a projected display using a mobile device such as a touchpad (Boring et al., 2010). However, the user has to memorize the video frame region surrounding the annotation, and then, once they shift focus to the real world, to find the reference point from memory. These focus shifts have been shown to be harmful in applications such as AR surgical telementoring, where it can lead to surgery delays or even errors (Andersen et al., 2016a).

User Perspective Rendering (UPR) attempts to improve the directional guidance by showing what the user should see if the display were not present. To achieve that, one needs to know the user head position as well as the scene geometry, such that the scene geometry can be reprojected to the user’s viewpoint (Babic et al., 2020).

The position of the user’s head was tracked with front facing cameras built into the device (Mohr et al., 2017). Researchers have also investigated not tracking the user head position and using instead a fixed default position (Copić Pucihar et al., 2013; Andersen et al., 2016b). We take the approach of not tracking the user head and of relying on the user to keep it at a relatively constant position with respect to the display.

A pure UPR approach has to know the geometry the scene to reproject the frame to the user’s view-...
point. The baseline provided by the two collaborators is insufficient for accurate triangulation of correspondences (Baricˇevi´c et al., 2017; Davison et al., 2007). Prior work has also proposed to do away with geometry acquisition and to work under the assumption that the scene is planar (Borsoi and Costa, 2018). The planar scene proxy is fitted in a calibration stage (Zhang et al., 2013), or tracked in real time from markers (Samini and Palmerius, 2014; Hill et al., 2011), or from scene features (Tomioka et al., 2013; S¨or¨os et al., 2011). If the scene is not close to the user, one can achieve a quality display transparency approximation under the "distant scene" assumption that ignores the distance between the device camera and the user viewpoint (Andersen et al., 2016b).

In our work we take the approach of not acquiring scene geometry. Instead, we rely on a homography computed with a prior-art computer vision pipeline that finds features, establishes correspondences between features, and minimizes correspondence reprojection errors (Bradski, 2000). Methods that acquire scene geometry passively, e.g., through monocular simultaneous localization and mapping (SLAM), have to "freeze" the visualizations on the collaborators’ displays in order to wait for the acquisition process to complete (Gauglitz et al., 2012; Bauer et al., 1999). The issue can be avoided by moving the scene acquisition to a preprocessing stage (Gauglitz et al., 2014). Our pipeline completes the initial transfer of frame features from the guide to the tourist in 300ms, and then computes the homography for each new frame in 20ms, without any preprocessing, which enables uninterrupted live visualization.

Co-located versus remote collaboration. When users are co-located, cross-referencing elements of a real world scene has the challenge that each of the users has its own view of the scene (Billinghurst et al., 2002; Kaufmann, 2003; Lin et al., 2015), and one has to transfer annotations from one view to the other. In remote collaboration, there is a single view of the real world scene which is transferred to the remote collaborator, who authors annotations directly in this shared view (Irlitti et al., 2016; Andersen et al., 2016a). We focus on co-located collaboration, and the annotation defined in the guide’s view is registered to the tourist view in real time.

3 LOOK-OVER-THERE

We have developed an AR approach for a collaborator (the guide), to point out an element of the real world (the reference point), to a second, co-located collaborator (the tourist), see pipeline in Fig. 2.

Stage A. The guide annotates the reference point on their device.
Stage B. The guide’s device extracts the salient frame features where the annotation was defined and transfers them wirelessly to the tourist device.
Stage C. The features are used by the tourist’s device to register the annotation, which is displayed where the reference point appears in the frame. Although the annotation does align with the reference point in the tourist tablet frame, the directional guidance provided is inaccurate. Indeed, the correct direction to the reference point is much lower and to the left, where the reference point appears in the field of view of the tourist (bottom left corner of C in Fig 2).
Stage D. The tourist aims their device such that the annotation is aligned with cross-hairs displayed at the center of the tablet. At the same time, the device zooms in to improve the simulated transparency accuracy. Finally, without changing view direction, the tourist removes the tablet to find the reference point directly with their naked eyes.

In summary, the guide creates the annotation using the touch-screen of their device (stage A). The guide frame where the annotation was authored is registered to the current (live) tourist frame using a homography mapping computed based on SIFT features (Lowe, 2004) (stage B and C). The robust SIFT algorithm allows us to not rely on the camera specs of the devices. The directional guidance refinement (stage D) is described in Sec. 3.1.

3.1 Directional Guidance Refinement

After the annotation registration stage, the annotation appears in the raw video frame of the tourist’s device. However, the directional guidance provided is inaccurate and a final refinement stage is needed.

Fig. 3 shows that before guidance refinement, the camera C of the tourist device captures the reference point T at image plane point Q. Q is shown on the display at point T’ which has the same pixel coordi-
Figure 3: Directional guidance accuracy improvement. The tourist viewpoint is $U$ and the reference point is $T$. Before guidance refinement, the tourist device points in the direction $UT'$ which has a large distance error $\varepsilon_d$ and a large angular error $\varepsilon_a$ with the true direction $UT$ to the reference point. Guidance refinement reduces both errors.

Figure 4: Annotation registration from guide to tourist.

4 EMPIRICAL EVALUATION

We have implemented our Look-over-there AR directional guidance method (Sec. 4.1) and tested it successfully on multiple scenes (Sec. 4.2). We have also evaluated our method in a user (Sec. 4.3).

4.1 Implementation Overview

We have implemented our method for Android devices using Android Studio. The timing data reported in this paper was recorded on a Samsung Galaxy Tab S6 tablet. We use OpenCV for the feature extraction and annotation registration stages of our pipeline (Fig. 2). When the guide creates an annotation, feature extraction finds between 20 and 100 salient SIFT (Lowe, 2004) features in the guide frame $F_{G0}$ where the annotation was created. A feature requires 568B, for a total of up to 50KB.

The features are transferred to the tourist device via Bluetooth. A homography is computed between the current tourist frame $F_{T_i}$ and the guide reference frame $F_{G0}$, and the homography is used to place the annotation on the tourist display. Since the tourist aims their device in the general direction of the reference point, and since the initial field of view of the tourist visualization is large, there is enough overlap between $F_{T_i}$ and $F_{G0}$ for the registration to succeed and for the reference point to be included in $F_{T_i}$.

Guidance refinement relies on the tourist to rotate the device such that the reference point be centered. The center of the display is marked with cross hairs for guidance. At the same time, the device zooms in, as described in Sec. 3. When the zoom in process is complete the annotation turns green.

4.2 Robustness

Our method works well on a variety of outdoor and indoor scenes, see Fig. 1, as well as the video accompanying this paper. Like in Fig. 1, the tourist’s direct view of the scene around their display is simulated, because it is hard to take a picture of the scene from the tourist’s viewpoint, and because in the outdoor scenes the display is dim. Fig. 4 illustrates the robustness of the annotation registration between the guide and the tourist device. Registration succeeds even when the reference point is close ($c$ and $d$).

4.3 User Study

We have conducted a controlled within-subjects user study ($N = 30$) with the approval of our Institutional Review Board. The goal of the study is to evaluate the directional guidance provided by our method.

Participants. We have recruited participants from the graduate and undergraduate student population of our university. The age range is 19 to 30, with an average of 25. 10 of the participants were women, 3 participants self-reported their prior experience with AR applications as "Never", 7 as "Once", 14 as "Occasionally", and 6 as "Frequently".

Conditions. The participant took the tourist role and the guide role was taken by a researcher. We compared the experimental Look-over-there condition (EC) to two control conditions.
dication (CCV), the guide provided verbal directions to the tourist to help them locate the reference point. The verbal directions were scripted and always the same. In a second control condition (CCA), the guide provided guidance to the tourist on a conventional AR display. Furthermore, we have also investigated a “blind mode” of Look-over-there, where the pipeline runs as usual, except that the tourist display shows the annotation over a white background, i.e., a blind experimental condition (BEC), as shown in Fig. 5. Since the frame is not shown, the tourist cannot memorize landmarks in the region of the annotation and has to rely exclusively on the directional guidance provided by the annotation. We do not advocate that BEC replace EC, BEC is just a condition used to measure the guidance provided by the annotation without the confounding factor of the visual context.

Experimental Setup. An essential requirement of our study is to be able to record objectively and accurately the location where the participant thinks the reference point is located after receiving guidance through one of the methods. We have considered two alternatives. One is to ask the participant to indicate verbally the location of the reference point. This option has the shortcomings that it depends on the participant’s verbal expression abilities, that verbal description is ambiguous, and that it cannot be used with CC as the participant can simply repeat the verbal instruction they were given. An alternative is to provide the participant a laser pointer to indicate the location. We have chosen this option. The challenge is that the laser dot is hard to see outdoors, and impossible to see when the reference point is beyond a few meters. For laser dot visibility and for condition repeatability, we ran the experiment in our laboratory. The scene is projected on a large screen.

Tasks and Procedures. We use one scene (Fig. 6a) in the training session and three scenes (b-d) in the actual tasks. The scenes differ in the availability and the visual salience of landmarks that can be used to locate nearby points of reference. The BlueDot scene (b) has a visually salient landmark, i.e., a single blue dot among orange dots. The Woods scene (c) has visual landmarks that are less salient, i.e., the trees are not identical but they are similar. The GreyDots scene lacks visually salient landmarks, i.e., all grey dots are identical. The dots in BlueDot and GreyDots are placed at irregular locations which prevents counting along cardinal directions. No reference point is reused (within or between conditions).

For BlueDot, the difficulty level is determined by the distance between the reference point and the landmark. The farther the reference point, the more difficult to describe it with verbal instructions (CCV). We stopped time when the participant indicated the target with the laser pointer. For a trial to be counted as successful, participants must locate the target (circle for BlueDot and GreyDots, and bird for Woods) on the first attempt. We measured the accuracy, i.e., percentage of successful trials, for GreyDots and Woods.

In CCV, the participant does not use the tablet. After the verbal instructions describing the location of the point of reference, the participant uses the laser pointer to indicate the location of the reference point. In CCA, the participant sees the annotation on their tablet and they shift their eyes back and forth between the tablet and the scene to locate the reference point in the scene. For Woods and GreyDots, the image on the tablet is a stationary image showing a sub-region of the scene that contains the annotation. This testing scenario matches the real world scenario and avoids the limitation of the experimental setup. When the participant thinks they found the reference point, they indicate its location in the scene with the laser pointer while looking at the scene. In EC and BEC, once the guidance is refined, the participant puts down the tablet without changing view direction and indicates the location of the reference point with the laser dot.
There is a single session per participant of up to 60min. A participant (1) fills in the demographics data form; (2) undergoes a 5-10min training session; (3) performs 3x3=9 trials in the BlueDot scene for three reference point locations and three conditions, i.e., CCV, CCA, and EC (the three locations for a condition had different difficulty levels, as shown in Fig. 6b); (4) performs 2x3=6 trials in the Woods scene for two reference point locations and three conditions, i.e., CCA, EC, and BEC; (5) performs 2x3=6 trials in the GreyDots scene like for the Woods scene.

**Data Collection.** The study collects the answers to the questionnaires, the time in seconds to locate each reference point, and whether the reference point was located correctly or not. The time interval measured starts from when the verbal instructions start for CCV and from when the annotation appears on the tourist tablet for the other conditions (CCA, EC, and BEC). The interval ends when the participant aims the laser pointer at the scene. The time is capped at 30s, after which, if a participant did not find the reference point, they move on to the next trial. For BEC we record the location indicated by the participant, i.e., the location of the laser dot instead of correctness.

**Data Analysis.** We performed a statistical analysis of the differences in task completion times for the three conditions and the three reference points of each scene. We used the Shapiro-Wilk test to check for data normality. When the normality assumption was verified we used a one-way repeated measures ANOVA test. When the normality assumption was not verified we used a Friedman’s non-parametric test. For significant three-level differences we performed a posthoc analysis of pairwise differences, with Bonferroni correction, either with a paired t-test (normal distribution), or with a Wilcoxon signed-rank test. We have also computed the proportions of correct reference point localizations for each condition. We investigated the statistical significance of the differences between conditions using McNemar’s test for paired proportions.

**Results and Discussion.** For the BlueDot scene, the completion time box plots are given in Fig. 7, for each of three conditions (CCV, CCA, and EC), and for each of three reference points. The reference points are shown in Fig. 6b. Fig. 8 shows that most of the data is not normally distributed, i.e., $p < 0.05$, which calls for the nonparametric Friedman test, for which the table reports the median (IQR) levels $L$, the test statistics value $\chi^2$, and the significance level $p$. For CCV, there is a significant difference between the three reference points. The post-hoc Wilcoxon signed-rank test reveals that all pairwise differences are significant, using the Bonferroni corrected significance level of $p < 0.017$. As expected, the CCV times increase significantly from Easy to Medium, and from Medium to Difficult, as longer and longer verbal explanations are needed. For CCA, the Friedman test does not report a significant difference, and for EC, the significance is at threshold level. We have also conducted the same analysis without outliers, with the same conclusions.

Fig. 9 shows that there are significant differences between the three conditions, for each of the three reference point difficulties. Post hoc pairwise analysis reveals that, with a Bonferroni correct significance threshold of $p < 0.017$, CCV times are significantly longer than CCA and EC times, with the exception of CCV-EC for the simplest task. Furthermore, CCA times are significantly shorter than EC. We conclude...
that for a scene with a highly salient landmark, AR has an advantage over verbal, and within AR, the guidance refinement stage is not necessary. A user can quickly see and memorize the position of the annotation with respect to the landmark, e.g., the single blue dot, and then find the reference point from memory. The reference point localization accuracy is 100% for CCA and EC and 88±16% for CCV.

For the Woods and GreyDots scenes, locating reference points based on verbal instructions takes too long and accuracy is too low, so CCV does not need to be run on these two scenes. The scenes were used to differentiate between a conventional AR display and Look-over-there. The completion time box plots are given in Fig. 10. All data was normally distributed so a paired t-test was used. Fig. 10 shows that all differences are significant; EC is faster than CCA for both Woods and GreyDots, and GreyDots is more challenging than Woods for both CCA and EC. EC times are shorter than CCA times even though EC zooms-in on the target gradually. For Woods, localization accuracy was 93±17% for both CCA and EC. For GreyDots, localization accuracy was 60±30% for CCA and 87±29% for EC. We analyzed the difference with McNemar’s test for paired proportions which confirmed the significance ($p < 0.001$) of the accuracy advantage of EC over CCA for the challenging GreyDots scene.

The shorter completion times and higher accuracy of EC over CCA proves the effectiveness advantage of the directional guidance provided by Look-over-there compared to conventional AR hypothesized by our theoretical analysis ($H_1$). Furthermore, since no user head tracking was used, our empirical results also confirm that Look-over-there is robust with user’s head deviating from the center of the display ($H_2$).

**Blind Localization.** Our pipeline can guide the participant towards the reference point using the annotation on a white background alone without any memorization of the visual context (Fig. 5). The localization error is defined as the distance from the center of the circle annotating the reference point to the location indicated by the user with the laser dot. By taking into account the viewing distance (about 4 meters), the localization error can be expressed in degrees. The average localization error is $0.30±3.11$° vertically and $3.3±2.94$° horizontally for Woods and $0.31±1.24$° vertically and $0.29±1.19$° horizontally for GreyDots. These small errors indicate that even in the total absence of visual landmarks, Look-over-there provides visual guidance that is accurate in an absolute sense, and not just sufficient to isolate one of a small number of candidate landmarks.

## 5 CONCLUSIONS AND FUTURE WORK

We have presented Look-over-there, an AR approach for a collaborator to convey a real world scene reference point to a second, co-located collaborator. For a scene with repeated patterns, e.g., even-sized dots arranged in a regular 2D grid, or for a scene without features, e.g., a white wall hiding a pipe that has to be repaired, the guidance refinement of Look-over-there significantly reduces the localization time compared to a conventional AR display.

Our method works for points of reference beyond the immediate vicinity of the collaborators, i.e., beyond 2m. There is no upper limit on the distance to the reference point, and that scenario is where directional guidance is most needed. Indeed, when the reference point is close, walking up to it, pointing in the direction of, and even touching the reference point are simple solutions that work well.

Our method was demonstrated in the context of two collaborators, a guide and a tourist, but the future work can extend it to a group of tourists. The only scalability bottleneck is that the annotation frame features have to be broadcast to all tourists, which could be addressed by tourists communicating the features to each other in a number of steps that is logarithmic with the number of tourists. Once a tourist device has the features, it registers the annotation to its live frame independently of the other tourists, without implications on the scalability with the tourist group size.

Our work provides a practical and robust method for improving the quality of the AR interface provided by handheld devices. Our goal is to help close the gap in directional guidance quality between video see-through handheld AR displays and optical see-through AR HMDs, to leverage the handheld AR displays’ form factor, ubiquity, and social acceptance advantages, paving the road towards the wide adoption of AR technology.
ACKNOWLEDGEMENTS

This material is based upon work supported in part by the National Science Foundation under Grants No. 2219842 and 2318657. We thank Shuqi Liao for her help with the experiments.

REFERENCES


