Efficient In-Situ Creation of Augmented Reality Tutorials

Alexander Plopski*, Varunyu Fuvattanasilp†, Jarkko Polvi‡,
Takafumi Taketomi§, Christian Sandor¶, and Hirokazu Kato‖
Graduate School of Information Science, Nara Institute of Science and Technology
Ikoma, Japan
Email: *plopski@is.naist.jp, †varunyu.fuvattanasilp.ut0@is.naist.jp, ‡jarkko-p@is.naist.jp,
§takafumi-t@is.naist.jp, ¶sandor@is.naist.jp, ‖kato@is.naist.jp

***Abstract*—With increasing complexity of system maintenance there is an increased need for efficient tutorials that support easy understanding of the individual steps and efficient visualization at the operation site. This can be achieved through augmented reality, where users observe computer generated 3D content that is spatially consistent with their surroundings. However, generating such tutorials is a tedious process, as they have to be prepared from scratch in a time consuming process. An intuitive interface that allows users to easily place annotations and models could help reduce the complexity of this task. In this paper, we discuss the design of an interface for efficient creation of 3D aligned annotations on a handheld device. We also show how our method could improve the collaboration between a local user and a remote expert in a remote support scenario.**

***Index Terms*—Training, Handheld Augmented Reality, Augmented Reality, Remote Assistance, Interaction, Annotation**

I. Introduction

With the increasing complexity and short life cycle of devices the use of printed explanations for their manufacturing, evaluation, and maintenance is becoming less and less viable. As an alternative, interactive guidelines can be used to provide step-by-step support. While it is relatively easy to generate such tutorials for digital devices, such as hand-held devices, their usability may be suboptimal because the user's view does not coincide with that of the presented explanation [12]. Augmented Reality (AR) can help address this problem by presenting virtual content that is accurately aligned with the surroundings [1]. By presenting a step-by-step tutorial in AR it is possible to reduce the mental demand, the number of errors, and consequently the time it takes to perform a task [13].

Creating such tutorials presents a big hurdle to their wide application in the industry and private households. Currently, all tutorials are prepared by hand, which requires an expert, as well as an accurate model of the device that will be processed. The expert then has to prepare easy-to-understand visualizations that are aligned with the model by hand. This has to be repeated for every step of the process, before deploying it to the user. As one can imagine, it is a very time-consuming and expensive process.

In this paper, we present our ongoing research to simplify and reduce the time that is required to create such tutorials. Our motivation stems from the large number of tutorial videos that are available online, for example on YouTube. Hereby, an expert is performing maintenance of a device. A user, who is watching this video, can then follow it step by step. We imagine that the expert could use the same process to create a tutorial that will be shared with the user by, for example, placing annotations that outline the next steps onto the device while maintaining it. Such in-situ editing has been used in [4] to let users create interactive AR games.

We believe that handheld devices can be efficiently used during the authoring process. Handheld devices have been used as an authoring tool in the past [7], [8]. They are widely available and are equipped with a variety of sensors, ranging from cameras to inertial measurement sensors that can be used to provide information about the devices state. A major challenge is how to efficiently place 3D content with a handheld device. As the device itself only presents a 2D interface, it is necessary to design methods that enable simple and efficient placement of annotations into the scene.

Users could adjust the pose of virtual objects with real and virtual buttons [2], [5]. However, this is a very cumbersome task. To simplify this process, Jung et al. [6] use single and multitouch gestures instead of buttons to position objects. Henrysson et al. [5] have also suggested that instead of using only gestures, a combination with the device movement could lead to superior results. Marzo et al. [9] combined the advantages of gesture manipulation and device movement methods to improve the speed at which users can place models.

Some methods take advantage of the device's sensors and the features of the environment, to estimate surfaces and prealign models according to the surface's normal [11]. However, such methods greatly depend on the accuracy of the reconstructed surface, are affected by noise, or cannot recover a suitable surface in complicated environments. Furthermore, this alignment may not correspond to the user's intention, and would require further adjustment.

In this paper, we present SlidAR, a method that allows users to efficiently place annotations into a scene. While in [5] the device movement was used as input for positioning of the virtual content, we use it primarily as a means to control the viewpoint, and adjust the position of the content with gestures on the display. We describe SlidAR+, an extension of SlidAR that lets users augment the scene not only with annotations, but also with 3D models. Finally, we discuss how our methods

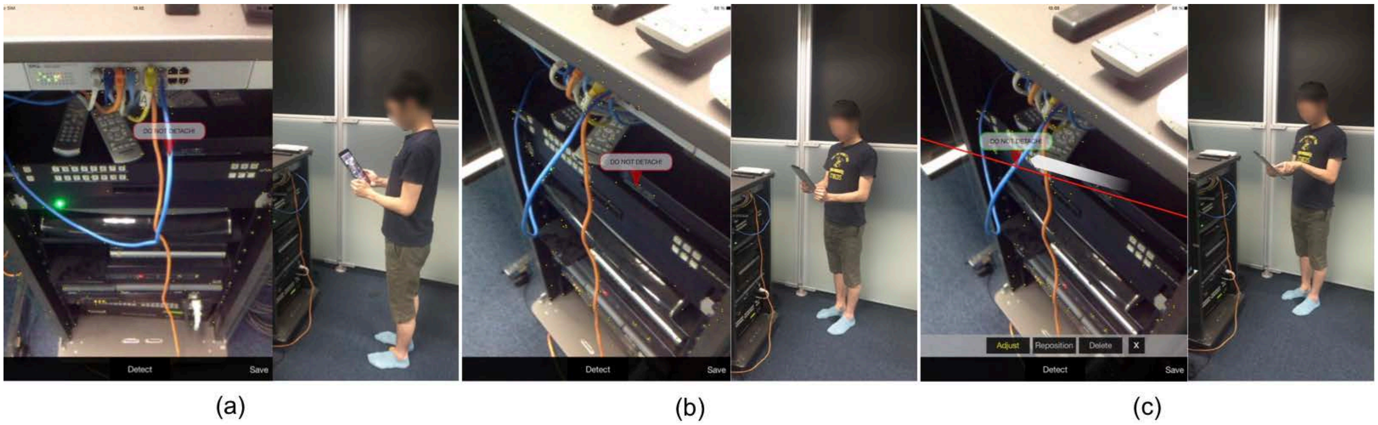

Fig. 1. Example of a user placing an annotation with SlidAR. (A) The user places the label "Do not detach" onto the blue cable. (b) After shifting the user's viewpoint, the label appears misplaced. (c) The user can adjust the position of the label by sliding it along the ray it was seen at from the previous position. (Figure taken from [15])

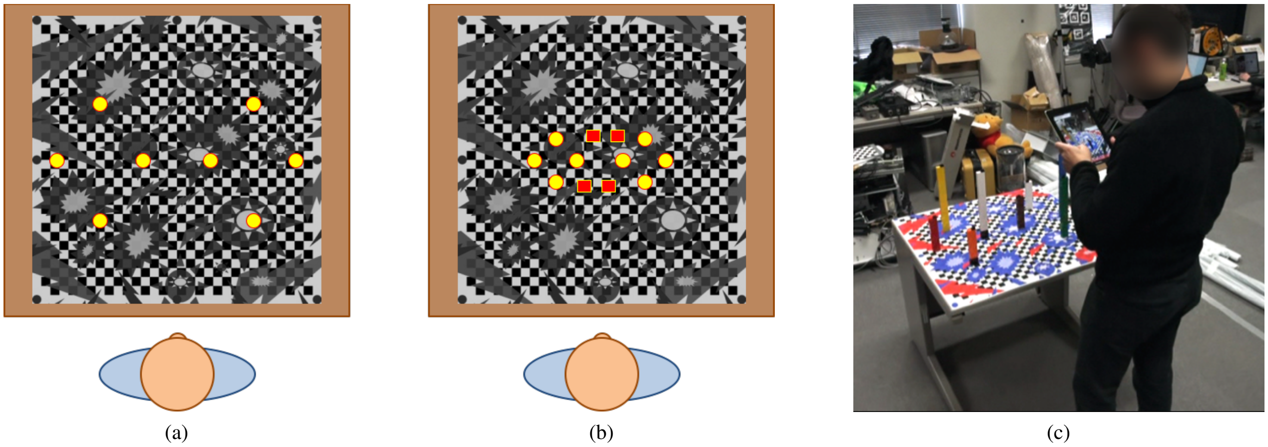

Fig. 2. We evaluated two scenarios, (a) an easy scenario where participants placed labels at 8 sparsely distributed locations (yellow circles) and (b) a difficult scenario that contained 8 densely placed target locations (yellow circles) and 4 distractors (red squares). (c) A user is placing a label in the easy scenario.

can also be applied to enhance real-time remote collaboration between multiple users.

II. Annotation Placement

The simplest way to provide guidance, is to present labels over the corresponding objects. One major concern when placing annotations into the scene, is how to properly position them. While it is possible to automatically align the annotations with the user's view, users must adjust 3 degrees of freedom (DoFs), namely the translation along the x, y, and z axes. Controlling all 3 DoFs is difficult and time-consuming. It may even lead to confusion, if the object behaves different from what the user expects because the systems coordinate system does not properly align with his current view. Our method separates this process, into an initialization phase that determines 2DoFs, and an alignment phase where the user only has to adjust 1 DoF. We call this method SlidAR and show the process in Fig. 1. During the initialization phase, the user can create an annotation and selects where it should appear from his current viewpoint. The annotation is initialized at a fixed distance along the ray cast from the pixel selected on the handheld display. After this step, the view is identical to what users would observe in a classic guideline. As the user shifts his viewpoint he will notice a misalignment of the label with the intended position. When the user wants to adjust the label's position, he sees a red line that represents the ray that the label was placed upon. Now, he can slide the label along this ray to the intended position.

We conducted a study, where we compared SlidAR with HoldAR that was introduced by Henrysson et al. [5]. HoldAR is a device movement based annotation technique. After initializing the annotation at a desired location, the user can perform a tap-and-hold gesture to fix the position of the label relative to the camera. Now, the user can adjust the label placement by moving the handheld device. To help users better understand where the model is in space, the label is casting a shadow directly below it onto the ground plane, and a red line connects the shadow and the label. Both SlidAR and HoldAR have to track the device's movement to allow placement, adjustment, and accurate presentation of the augmentations. We track the device's position through Simultaneous Localization and Mapping (SLAM). SLAM algorithms predict the camera motion by tracking how features shift between consecutive frames [17]. By generating a map of keyframes, these algorithms can also recover from tracking failure by matching the current frame to the collected keyframes. To keep the experiment conditions the same, we used the same pre-generated feature maps for both methods, and disabled the generation of new feature points.

For our study, we recruited 23 graduate students (16 male and 7 female; mean age 29±5 years; age range 22 to 41; mean height, 167.5±12.8 cm), and asked them to place labels on top of Lego blocks placed at pre-defined positions with SlidAR and HoldAR. We evaluated the performance of the methods in an easy and a difficult scenario, as shown in Fig. 2. In the easy scenario (Fig. 2a) the environment contained 8 sparsely distributed Lego blocks and participants had to place a label on top of each block. In the difficult scenario (Fig. 2b), the 8 target locations were placed closely to each other. Additionally, we placed 4 distractor Lego blocks between the target locations.

We compare the methods based on the performance time, the magnitude of the misalignment with the intended position, and the average amount of device movement needed for this task. We found that SlidAR was significantly faster than HoldAR ($F(1, 22) = 28.08, p < .001, p.e.s. = 0.56$), and required less device movement ($F(1, 22) = 31.47, p < .001, p.e.s. = 0.59$) for both scenarios. Participants also reported that SlidAR was easier to use and to understand than HoldAR. We have presented a detailed description of the conducted experiment and its results in [14].

III. Model Placement

One major limitation of the current system is that it allows users to only place labels, which are less expressive than 3D models. Placing 3D models however, requires the user to be able to easily manipulate 7 DoFs (3 rotation, 3 translation, 1 scale), while the current system only supports the manipulation of the translation. Our current research focuses on the development of intuitive ways to place and adjust the rotation and scale of models. The main speed-up of SlidAR compared to previous methods is the constraint of the DoFs the user has to manipulate. By constraining the DoFs users have to manipulate during the rotational alignment we similarly expect a simplification of the alignment process. This will lead to an improved accuracy and reduce the time required for this process.

SLAM based systems initialize their coordinate system relative to the initial pose of the device, which results in a random orientation of the virtual content when it is placed into the scene. However, most man-made structures in our surroundings have either horizontal or vertical surfaces. In most cases, it is therefore sufficient to align the model parallel or perpendicular to the gravity vector. In the ideal case, after the pre-alignment the user will have to manipulate to only 1 rotational DoF. Most state-of-the-art head-mounted and hand-held devices are equipped with gyroscope sensors that provide the gravity direction at any given moment. We exploit this to automatically align the virtual content users place into the scene, independent of the orientation of the tracking component's coordinate system. An example of a user placing and adjusting the orientation of a 3D model with our system is shown in Fig. 3. In some cases, users may want to orient the model neither horizontally, nor vertically. To allow users to control all 3 rotational DoF, we implemented two-finger twist gesture to perform rotation around the z-axis (Z-Rot) [9] and ARCBALL [16] vertical slide gesture for y-axis rotation. Both of these functions rotate the object based on the current perspective. Furthermore, users can scale the object with a simple pinch gesture. We refer to SlidAR extended with capabilities to control all 7 DoF as SlidAR+.

We compare SlidAR+ with Hybrid, a state-of-the-art method that was shown to perform better than a device-movement method like HoldAR. Hybrid was introduced by Marzo et al. [9] and combines device-movement and screen-based manipulation. Hybrid, takes advantage of the user's capability to rapidly move the device to adjust the position, and fine control on the display to control the rotation.

Our preliminary experiments show that our method performs faster and requires less device movement than Hybrid for placing and orienting models in the scene, when these are aligned with, or perpendicular to the gravity direction. Our next goal is to perform a formal study where we also investigate how SLidAR+ performs for scenarios where the intended orientation is independent of the gravity direction.

IV. Remote Collaboration

The techniques presented in this paper can also be applied to support remote collaboration. A common scenario is a remote expert who helps a local user interpret the information provided by various sensors and to perform the correct steps during maintenance. In the past years, several applications for handheld devices [3] as well as head-mounted displays [10] have become more common. To facilitate efficient collaboration it is necessary to enable the remote partners to efficiently exchange information. This information should be also placed into the world as 3D objects, to ensure that the perceived guidance is not affected by a shift in the user's viewpoint.

We believe that SlidAR and SlidAR+ can be applied in this scenario to facilitate easy placement and adjustment of annotations for AR guidance on hand-held and head-mounted displays. By sharing the local user's view (the image captured by the camera of the handheld device, or the head-mounted display), the remote user can place labels and models that match the current viewpoint of the local user. After the local user shifts his viewpoint, these annotations are likely to appear at an incorrect depth. As the remote user will become aware of this, he can use SlidAR or SlidAR+ to adjust the positioning of these annotations. When using SlidAR+ to pre-align the orientation of the models, the system can take advantage of the local sensors to place it correctly in the local user's environment.

While there are a number of methods that allow to present and visualize annotations between remote users, such as projector based systems, using our approach presents a series of

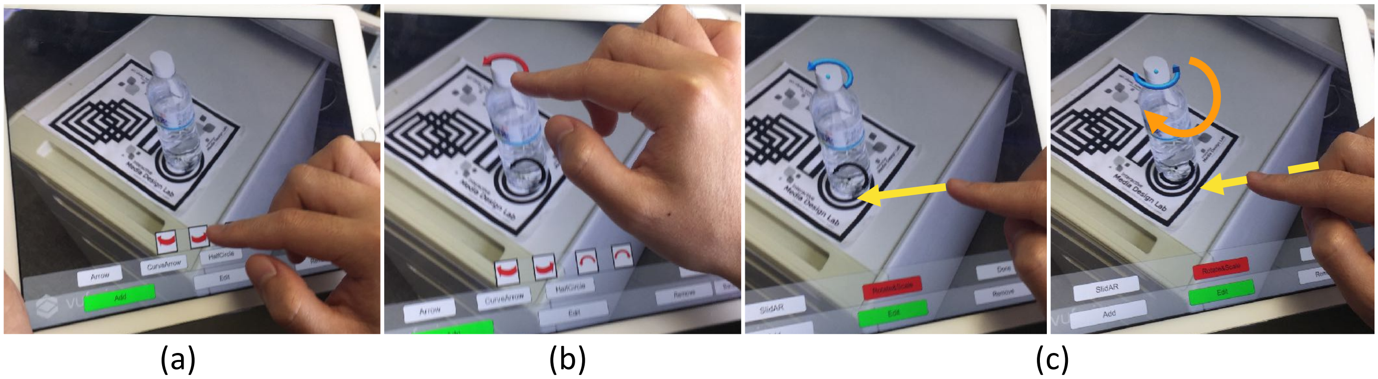

Fig. 3. Example of a user placing a 3D model into the scene. (a) After selecting the desired model, (b) the user positions it in the scene with SlidAR. (c) By swiping on the display the user can rotate the model around the gravity vector.

benefits. For one we do not require sophisticated devices and extended setup procedures. There is no need, to ensure that the remote environment matches that of the local user, as we do not share the 3D model between the users. All communication is based on the images and only the virtual content is placed in a 3D context. The remote user, can thus see the same augmented view as the local user. As the users share the same view, the remote user can easily spot potential errors, which helps align the mental states of the users. This also removes the need to track the remote user's viewpoint, as all augmentation is based on the local user's view.

Our system can also be used to allow the local user to share annotations and labels with the remote user. For example, a local user who uses a head-mounted display, can use SlidAR and SlidAR+ to place labels and models on a handheld device, or to adjust the pose of already placed models. By synchronizing the pose of the devices, these models would be visible to the remote user as well. Such two-way manipulation could further support the communication and assist the collaboration. In the future, we plan to conduct a formal study to evaluate how SlidAR and SlidAR+ affect remote collaboration.

V. Conclusions

In this paper we present two methods for placement of labels and models for intuitive generation of tutorials for AR maintenance. Our systems reduce the mental demand and time required to create the tutorials be reducing the number of DoFs users have to control to correctly place and align the augmentations. In the ideal case, our system will allow users to place models into the scene, by manipulating only 3 DoFs (1 translational, 1 rotational, and 1 scale). Further studies are necessary to inspect how well the orientation adjustment performs in the case where our assumption does not hold.

One major drawback of SlidAR and SlidAR+ is that both methods require very accurate initial placement of the model on the display, as this direction is used to adjust the position of the label. One of our future goals is to enable users to adjust erroneous initialization placements. For example, users could freeze frames to adjust the position of the label on the display.

We believe our system to be applicable not only for in-situ authoring, but also remote collaboration. Because the remote expert observes the same view as the local user, he can, therefore, easily detect and correct misalignments or incorrect placements. In the future, we plan to conduct a formal user study that compares SlidAR and SlidAR+ with existing methods in the remote collaboration scenario.

References

[1] R. T. Azuma. A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, 1997.

[2] R. Castle, G. Klein, and D. W. Murray. Video-Rate Localization in Multiple Maps for Wearable Augmented Reality. In *Proceedings of the IEEE International Symposium on Wearable Computers*, pages 15–22, 2008.

[3] S. Gauglitz, B. Nuernberger, M. Turk, and T. Höllerer. In Touch with the Remote World: Remote Collaboration with Augmented Reality Drawings and Virtual Navigation. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, pages 197–205, 2014.

[4] N. Hagbi, R. Grasset, O. Bergig, M. Billinghurst, and J. El-Sana. In-Place Sketching for Content Authoring in Augmented Reality Games. In *Proceedings of the IEEE Virtual Reality Conference*, pages 91–94, 2010.

[5] A. Henrysson, M. Billinghurst, and M. Ollila. Virtual Object Manipulation Using a Mobile Phone. In *Proceedings of the International Conference on Augmented Tele-existence*, pages 164–171. ACM, 2005.

[6] J. Jung, J. Hong, S. Park, and H. S. Yang. Smartphone as an Augmented Reality Authoring Tool via Multi-Touch based 3D Interaction Method. In *Proceedings of the ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 17–20, 2012.

[7] S. Kasahara, V. Heun, A. S. Lee, and H. Ishii. Second Surface: Multi-User Spatial Collaboration System based on Augmented Reality. In *SIGGRAPH Asia 2012 Emerging Technologies*, pages 20:1–20:4, 2012.

[8] T. Langlotz, S. Mooslechner, S. Zollmann, C. Degendorfer, G. Reitmayr, and D. Schmalstieg. Sketching Up the World: In Situ Authoring for Mobile Augmented Reality. *Personal and Ubiquitous Computing*, 16(6):623–630, 2012.

[9] A. Marzo, B. Bossavit, and M. Hachet. Combining Multi-touch Input and Device Movement for 3D Manipulations in Mobile Augmented Reality Environments. In *Proceedings of the ACM Symposium on Spatial User Interaction*, pages 13–16, 2014.

[10] J. Müller, R. Rädle, and H. Reiterer. Remote Collaboration With Mixed Reality Displays: How Shared Virtual Landmarks Facilitate Spatial Referencing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 6481–6486, 2017.

[11] B. Nuernberger, E. Ofek, H. Benko, and A. D. Wilson. Snaptoreality: Aligning Augmented Reality to the Real World. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1233–1244, 2016.

[12] S. Pathirathna, C. Sandor, T. Taketomi, A. Plopski, and H. Kato. [Poster] Video Guides on Head-Mounted Displays: The Effect of Misalignments on Manual Task Performance. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, pages 9–10, December 2016.

[13] J. Polvi, T. Taketomi, A. Moteki, T. Yoshitake, T. Fukuoka, G. Yamamoto, C. Sandor, and H. Kato. Handheld Guides in Inspection Tasks: Augmented Reality vs. Picture. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[14] J. Polvi, T. Taketomi, G. Yamamoto, A. Dey, C. Sandor, and H. Kato. SlidAR: A 3D Positioning Method for SLAM-based Handheld Augmented Reality. *International Journal of Computers and Graphics*, 55:33–43, December 2015.

[15] J. Polvi, T. Taketomi, G. Yamamoto, C. Sandor, and H. Kato. [DEMO] SlidAR: A 3D Positioning Technique for Handheld Augmented Reality, October 2015.

[16] K. Shoemake. III.1. - Arcball Rotation Control. In P. S. Heckbert, editor, *Graphics Gems*, pages 175 – 192. Academic Press, 1994.

[17] T. Taketomi, H. Uchiyama, and S. Ikeda. Visual-SLAM Algorithms: A Survey from 2010 to 2016. *ISPJ Transactions on Computer Vision and Applications*, 9(1):16:1–16:11, 2017.