AR Replay in a Small Workspace

Yun Li * Yoshinari Kameda † Yuichi Ohta ‡

University of Tsukuba, Japan

ABSTRACT

We propose "AR replay", a framework to record the working scene including a tutor's action in a small workspace, and then replay the tutor's action in front of a learner's view in an AR fashion (Figure.1). This framework uses one RGB-D camera for recording and replaying.

On learning a task in a small workspace, when a tutor cannot be in the workspace, it is useful for a learner to check the action of the tutor by a video which was taken in advance in the same workspace. If the video can be replayed in an AR fashion, it will be more useful. We propose a new "AR replay" method by using one RGB-D camera. In our "AR replay", the action of tutor is aligned to the right place and the learner can check the action from various viewpoints. The action is shown as 3D dynamic shape with color and it is aligned to the workspace by the static geometric clues in the workspace.

Since we expect the RGB-D camera is maneuvered to frame the interaction between the tutor and the static workspace environment, we assume the demand of changing viewpoint from the original recorded camera viewpoint is limited to some extent on checking the "AR replay".

Our preliminary experimental system can acquire the 3D shapes about tutor's action and the workspace environment. Moreover, this system can produce the "AR replay" on a video see-through display, with which a learner can shift the viewpoint from the original path of the RGB-D camera in order to have the better view of the interaction between the tutor and the static workspace environment.

Figure 1: Concept of "AR replay".

* email: s1320898@u.tsukuba.ac.jp

† email: kameda@iit.tsukuba.ac.jp

‡ email: ohta@iit.tsukuba.ac.jp

Keywords: AR replay, Camera tracking, 3D video, 3D shape reconstruction, RGB-D camera, Kinect.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems - Video; I.3.3 [Computer Graphics] Picture/Image Generation - Digitizing and Scanning.

1 INTRODUCTION

On learning a task in a small workspace, AR based systems have been proposed to support a learner to work by overlaying some visual instruction information about the task. For example, [9] provides a prototype AR support system on a procedural task and [10] provides a guidance for maintenance operations using automatic markerless AR system. In most cases, 3D CG objects are used to indicate visual instruction by analyzing the task so that they must be customized to each task. However, the customization could be complex and difficult for ordinary users.

Conventionally, it is considered to be useful for a learner to check the action of a tutor by a video which was taken in advance in the same workspace. If the tutorial video can be replayed in an AR fashion, it will be more useful. [11] proposes an AR support system by using existing 2D videos as instructional videos and overlaying the video onto the user's view space. However, the user just watches the video from recorded viewpoint since the video is recorded in 2D. This system is limited to the tasks on a desk and it needs a point marker to compute the position.

We propose a new "AR replay", a framework to record the working scene including tutor's action and workspace, and replay it in front of the learner by registering the recorded action of the tutor to the static workspace environment. In this paper, we define that the whole working scene in the workspace consists of dynamic tutor's action and static workspace environment. Some objects moved by the tutor are also classed to the dynamic tutor's action. The dynamic tutor's action will be acquired and replayed as a point cloud representation. The static workspace environment will be acquired as volume form and it is also used to align the pre-recorded 3D shape in the workspace. On recording the working scene, one RGB-D camera is used. On replaying, the learner has the RGB-D camera in his/her hand and he/she watches the replay of the dynamic tutor's action on the real workspace in video see-through fashion.

Our proposed method includes two main steps. The first step named "working-scene record" consists of (1a) recording the dynamic tutor's action and the static workspace environment, (1b) rebuilding the static workspace environment volume in 3D, and (1c) segmenting the dynamic action point cloud from the working scene. The second step named "AR view" consists of (2a) estimating the pose of RGB-D camera over the static workspace environment, (2b) aligning and displaying the dynamic action of the tutor on the video see-through display (Figure 2).

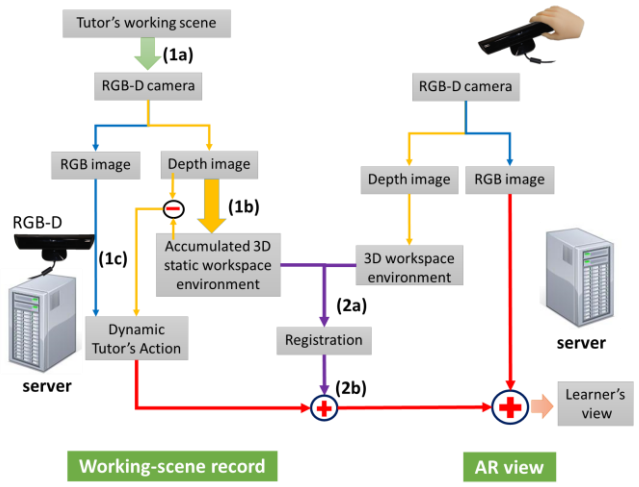

Figure 2: Block diagram of "AR replay".

2 Related Work

As for the working-scene recording step, if the whole working scene including both the tutor's dynamic 3D shape and the static workspace environment could be captured and reconstructed on-line, it will be perfect, but it is not an easy mission. Fully automated approach such as [3] and rapid building approach with a little help of human hand [4] have been proposed, but they cannot record and store the dynamic tutor's action because they focus on static objects. A large scale change over timeline can be visualized by [5], but they are not focusing on continuous dynamic actions. The data from multi-camera is integrated in order to acquire a fully 3D scene including both the dynamic action and the static environment. Usually, billboard method is used to record a large scene such as a soccer match [17] and volume intersection is used to record a stage performance [6][15][16]. However, this method is not suitable for a working scene, since they need a number of cameras and it is not easy to find appropriate camera positions in a small workspace.

Recently, approximate 3D video could be acquired just by using a small quantity of sensor camera. Such as encumbrance-free telepresence system offers fully dynamic, real-time 3D scene capture and continuous-viewpoint which is based on an array of commodity RGB-D cameras [7][8][12]. This capture system achieves a good quality result. A remote user could look around the scene from various viewpoints. Even though the number of RGB-D camera needed in this approach is small, the range of the space which can be recorded is limited since the cameras are fixed.

In a workspace, usually, available space for camera setting is limited because some machines and devices have been located, therefore, it is more practical to record the working scene by adjusting the viewpoint of camera rather than setting multi cameras in the workspace. As a reconstruction system just by using one handheld RGB-D camera, KinectFusion [1][2] creates high-quality, geometrically accurate, 3D models in real-time by integrating live depth data into a global volume. Therefore, complex setting and camera calibration is not required in this method. [1] and [2] focus on scanning and reconstructing static object or room space by a single volume. This method also shows the possibility of reconstructing dynamic object such as human by using a second volume. However, it is difficult to keep record and store the dynamic motion continually because the volume data is enormous.

Our "AR replay" system records a working scene just by using one RGB-D camera based on extending KinetFusion algorithm. We assume that a handheld camera should be properly maneuvered to frame important interactions between the tutor and the static workspace environment. It results in the assumption that learner may want to shift his/her viewpoint on "AR replay" slightly from the original camera position in order to have a better view of the interaction, or to step-in/out the scene for better understanding of tutor's action in the workspace.

3 Working-Scene Recording

In this chapter, we describe the acquisition process of the static workspace environment volume and the dynamic tutor's action point cloud in the working scene.

Our "AR replay" system apply a learner could check the pre-recorded 3D shape of tutor's action in a working scene taken by a RGB-D camera in the same workspace. On working-scene record step, it is necessary to separate and record a dynamic tutor's action and a static workspace environment from the working scene. We use and extend KinectFusion [1][2] to achieve it. Then, the dynamic tutor's action will be acquired and replayed as point cloud representation. The static workspace environment will be acquired as volume form and it is also used to align the pre-recorded 3D shape in the workspace on AR view step.

Our "AR replay" uses KinectFusion [1][2] to rebuild the static workspace environment. KinectFusion is proposed to reconstruct an indoor scene and register the camera pose in the scene by fusing all of the depth data streamed from a Kinect sensor (RGB-D camera) into a single 3D global model of the observed scene in real-time. As KinectFusion rebuilds only geometric shape of the static workspace environment, we develop the functions to attach color attribute to the 3D global model and to save both the geometry and shape data. The saved static workspace is represented as colored 3D point clouds.

The recording should start before the tutor comes into the workspace. The RGB-D camera is moved to glance the small workspace. It helps learners to understand the layout of the workspace, and it is also good for Kinect-Fusion to realize rapid building of initial 3D global volume model. Then, the camera should be maneuvered to frame the action of the tutor over the static objects in the workspace. While the camera is framing the action, 3D shape of the static workspace environment is automatically updated.

There are 4 sub-steps to acquire the dynamic action of the tutor and the static workspace environment (Figure 3). Yellow lines in Figure 3 indicate the data flow of depth/geometric data and blue lines indicate that of color data.

Sub-step A: Input acquisition. The RGB-D camera takes a depth image and an RGB image. The depth image is converted to 3D point clouds.

Sub-step B: Camera registration and scene separation is made by Kinect-Fusion. Currently observed surface obtained from the depth image is compared with the surface of the 3D global volume model, and the pose of the RGB-D camera is estimated. The points that are consistent to the surface of the 3D global model are marked as workspace, and the other points are marked as tutor's action.

Sub-step C: The points marked workspace is used to improve the global 3D volume model, which will become to cover wider area of the static workspace environment in better accuracy. Associated color attribute is also saved as accumulated workspace environment.

Sub-step D: The tutor's action is colorized by referring input RGB image. The action calibration parameters to the static workspace are also saved for the alignment in AR view stage.

When the action of the tutor is not needed to be placed on the real workspace in AR fashion, a learner can see the replay the action fully virtually by recalling both the saved action and the saved static workspace. Note that the reconstruction of the static workspace is widely made because it is a result of accumulation of the depth video.

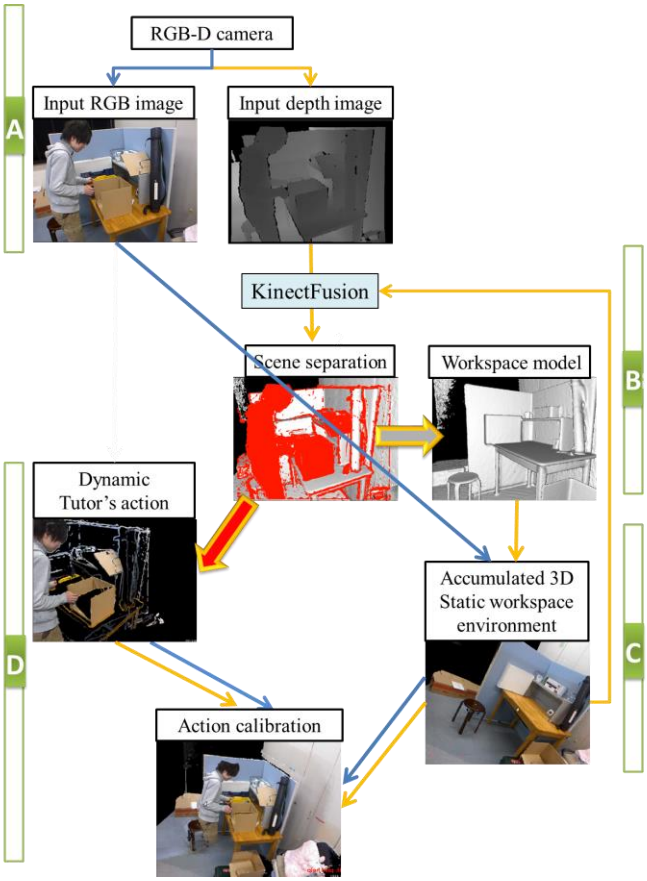

Figure 3: Data flow about working-scene record.

4 AR VIEW

On AR view, a learner has a RGB-D camera in his/her hand and see the recorded action of the tutor in video see-through display. Though the learner can move the camera freely, we ask the learner to start the replay from the recorded camera position at the beginning because following the original camera gives good view of the action basically.

The camera registration is done again by the same KinectFusion. The difference is that it starts with the saved static workspace environment as the 3D global model. Since the static environment is same, the estimated camera parameter to the static workspace environment can be used both for registering the camera to the real world and aligning the pre-recorded static workspace environment to the current real world.

Then, the replay of the action of the tutor is aligned and placed over the real workspace in video see-through display.

5 EXPERIMENTAL RESULT

5.1 Hardware Configuration

We use a single PC with 4-core Intel Core i7-3770 CPU, 4GB of RAM and Nvidia GeForce GTX660Ti graphics board for both working-scene record step and AR view step in our "AR replay" system. One Microsoft Kinect sensor is connected to the PC.

5.2 Result of Experiment

We have implemented preliminary "AR replay" system.

During the working-scene record step, we acquired both the point cloud of dynamic tutor's action and the volume of static workspace environment. Figure 4 shows a result of the working-scene record step. The upper image is the volume of accumulated 3D static workspace environment with color. The lower image is a snapshot of the point cloud of dynamic tutor's action. The point cloud of dynamic tutor's action on each frame is combined into a point cloud stream. Moreover, the volume of workspace environment is saved with compression for reducing the storage space. Our extended KinectFusion algorithm took 45 ms and point cloud extraction took 38 ms on average. The initial volume resolution is 512^3, compressed volume in a size of 2~5 MB and took about 10 s.

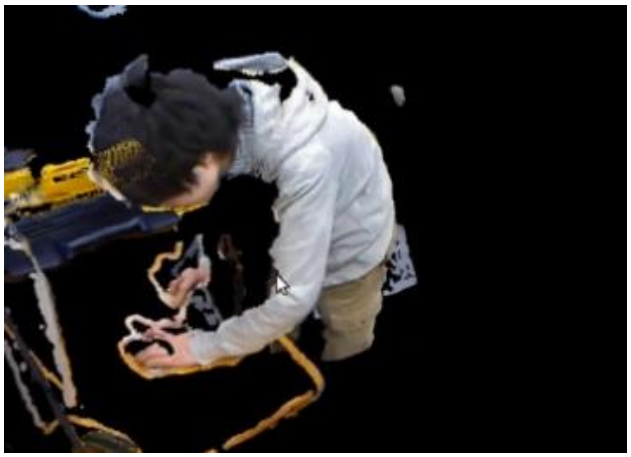

Figure 4: Result of working-scene record: upper) volume of static workspace environment, lower) point cloud of tutor's action.

The upper pictures of Figure 5 are the original input RGB images (A) at t_1 and t_2. The corresponding recorded working-scene is shown in lower pictures (B). Since a learner moves the viewpoint on watching the recorded tutor's working scene, we set the viewpoint close to the original camera position and took the snapshot of the virtual replay as B-t_1, and stepped back and took B-t_2.

The snapshot of B-t_1 indicates the replayed content of the working scene can convey almost same quality of the original image. It means that recorded working-scene have sufficient resolution and information for a learner to observe the tutor's action and the interaction with the workspace. The snapshot of B-t_2 shows that the accumulated workspace covers the wider area of the scene that is not visible in the original camera position. It enables a learner to shift the viewpoint slightly to have the better view of the interaction, or to step-in/out the scene in the workspace.

The "AR replay" has some minor limitations. The tutor's action that is not visible by the original RGB-D camera cannot be recorded and visualized. For example, the head at t_1 and legs at t_2 are missing in B-t_2. The other limitation is that the workspace at far range cannot be integrated because the RGB-D sensor cannot measure far places.

Currently we are developing the AR view stage. A preliminary snapshot of "AR replay" under progress is shown in Figure 6. The accumulated static workspace environment is compressed and saved as volume data (including vectors and weights). The tutor's action data is saved as point cloud stream. We have already developed a loading function of the saved static workspace model with compression and a matching function to align the model with real workspace environment. Figure 6 is a result of the matching function. The saved static workspace model can be observed from the current viewpoint of a learner.

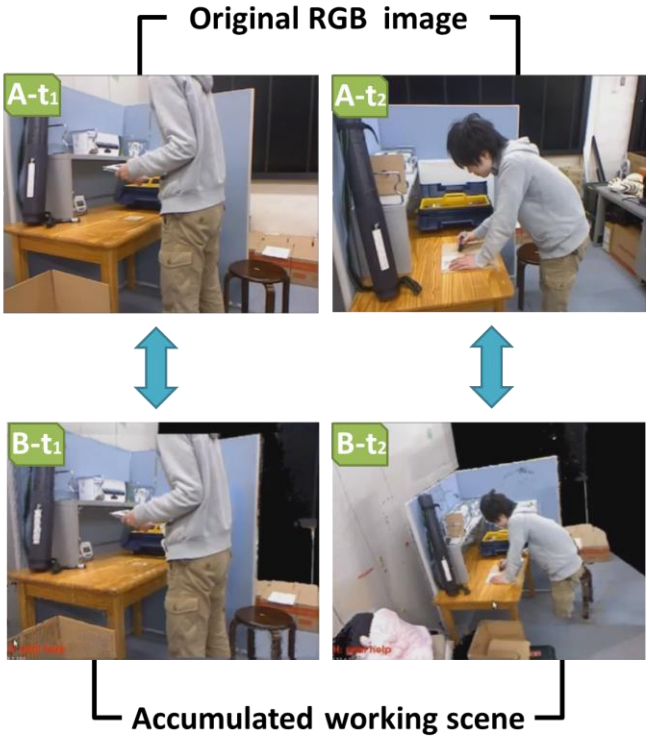

Figure 5: Two snapshots.

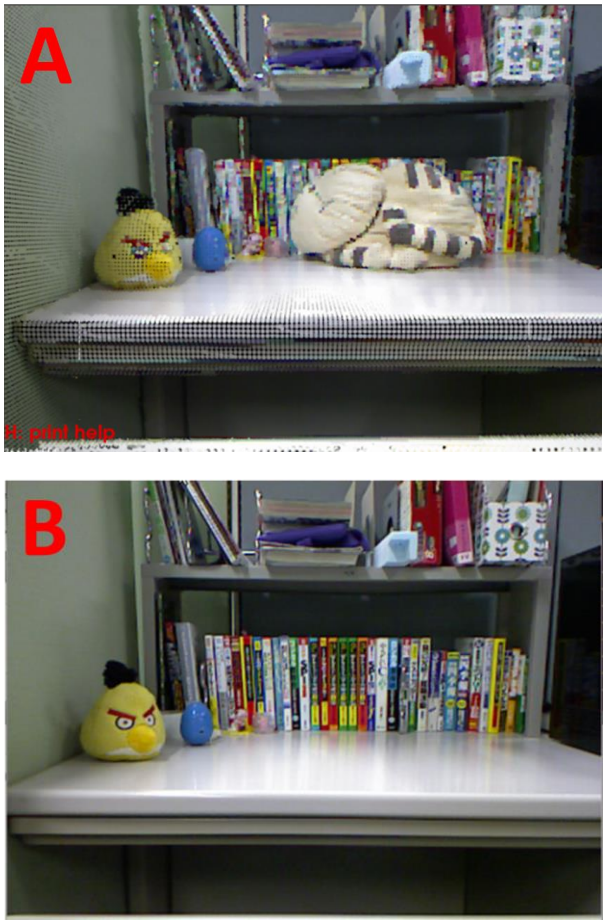

Figure 6: Loading and matching of the workspace
(A): saved static workspace model (B): current view of leaner.

6 Conclusion and Future Work

We propose a novel "AR replay", a framework to record the working scene including the action of a tutor and the workspace around the tutor, and replay it in a view of a learner by registering the recorded 3D tutor's action into the real workspace. On recording the tutor's action, only one RGB-D camera is used. The "AR replay" let the learner watch the tutor's action in the workspace in a video see-through display with a RGB-D camera. This method includes two main steps. One is how to acquire a tutor's action and the workspace environment during the recording of a working scene. The other is how to align the pre-recorded tutor's action at the real workspace.

As for the first step of working-scene record, we have presented a method for acquiring tutor's action and the accumulated workspace model by integrating depth data of a scene based on KinectFusion algorithm.

As for the second step of AR view, we successfully loading the saved static workspace environment model and matching the model with the real workspace environment. However, full functionality of the "AR replay" system has not been realized yet because of some troubles. This research is working in progress.

There are three features of the "AR replay". First, accumulated static workspace is created in real-time. It needs only one RGB-D camera. Second, the 3D workspace model is reconstructed automatically during the recording process while the camera is placed to frame the tutor's actions. Finally, the learners are

capable of grasping the video from different viewpoints in permissible level, although it was captured from single viewpoint.

This research is ongoing work and the complete "AR replay" system with full functionality will be presented in near future.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 23300064.

REFERENCES

[1] R. A. Newcombe, S. Izadi et al. KinectFusion: Real-time Dense Surface Mapping and Tracking. International Symposium on Mixed and Augmented Reality (ISMAR), pp. 127-136. 2011.

[2] S. Izadi, D. Kim, O. Hilliges, R. Newcombe, A. Fitzgibbon, et al. KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. The Symposium on User Interface Software and Technology (UIST), pp. 559-568, 2011.

[3] Y. Furukawa, B. Curless, S. M. Seitz, R. Szeliski, Reconstructing Building Interiors from Images. International Conference on Computer Vision (ICCV), pp. 80-87, 2009.

[4] T. Ishikawa, T. Kalaivani, M. Kourogi, A.P. Gee, W. Mayol, K. Jung, T. Kurata. In-Situ 3D Indoor Modeler with a Camera and Self-Contained Sensors. Virtual and Mixed Reality (HCII2009), LNCS 5622, pp. 454-464, 2009.

[5] S. Zollmann, D. Kalkofen, C. Hoppe, S. Kluckner, H. Bischof, G. Reitmayr. Interactive 4D Overview and Detail Visualization in Augmented Reality. International Symposium on Mixed and Augmented Reality (ISMAR), pp. 167-176, 2012.

[6] W. Matusik, C. Buehler, R. Rasker, S.J. Gortler, L .McMillan. Image-Based Visual Hull, SIGGRAPH'00, pp. 369-374, 2000.

[7] A. Maimone, H. Fuchs. Encumbrance-Free Telepresence System with Real-Time 3D Capture and Display using Commodity Depth Cameras. International Symposium on Mixed and Augmented Reality (ISMAR), pp. 137-146, 2011.

[8] A. Maimone, H. Fuchs. Real-time volumetric 3D capture of room-sized scenes for telepresence. The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), pp. 1-4, 2012.

[9] J. Henderson, K. Feiner, Augmented Reality in the Psychomotor Phase of a Procedural Task. International Symposium on Mixed and Augmented Reality (ISMAR), pp. 191-200, 2011.

[10] H. Alvarez, I. Aguinaga, D. Borro. Providing Guidance for Maintenance Operations Using Automatic Markerless Augmented Reality System. International Symposium on Mixed and Augmented Reality (ISMAR), pp. 181-190, 2011.

[11] M. Goto, Y. Uematsu et al. Task support system by displaying instructional video onto AR workspace. International Symposium on Mixed and Augmented Reality (ISMAR), pp. 83-90, 2010.

[12] A. Maimone, H. Fuchs. A First Look at a Telepresence System with Room-Sized Real-Time 3D Capture and Life-Sized Tracked Display Wall. International Conference on Artificial Reality and Telexistence (ICAT), pp. 67-72, 2011.

[13] S. Rusinkiewicz , M. Levoy. Efficient Variants of the ICP Algorithm. Third International Conference on 3D Digital Imaging and Modeling, pp. 145-152, 2001.

[14] P. Besl, N. McKay. A Method for Registration of 3D Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 14(2), pp. 239-256, 1992.

[15] T. Kanade, P. Rander, P. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. IEEE Multimedia, Immersive Telepresence, 4(1), pp. 34-47, 1997.

[16] T. Matsuyama, X. jun Wu, T. Takai, S. Nobuhara. Real-Time 3D Shape Reconstruction, Dynamic 3D Mesh Deformation, and High Fidelity Visualization for 3D Video. Computer Vision and Image Understanding, 96(3), pp. 393-434, 2004.

[17] T.Koyama, I.Kitahara. Y.Ohta. Live Mixed-Reality 3D Video in Soccer Stadium. International Symposium on Mixed and Augmented Reality (ISMAR), pp. 178-186, 2003.