TutAR: Augmented Reality Tutorials for Hands-Only Procedures

Daniel Eckhoff
Nara Institute of Science and Technology
Ikoma, Japan
daniel.eckhoff.dz3@is.naist.jp

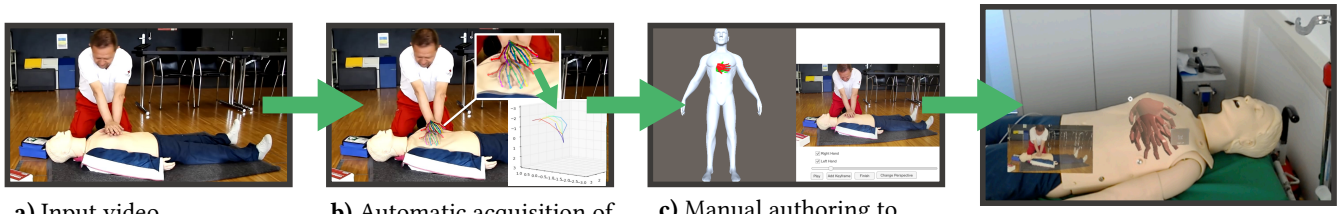
Christian Sandor
Nara Institute of Science and Technology
Ikoma, Japan
sandor@is.naist.jp

Christian Lins
OFFIS
Oldenburg, Germany
christian.lins@offis.de

Ulrich Eck
Technical University of Munich
Munich, Germany
ulrich.eck@tum.de

Denis Kalkofen
Graz University of Technology
Graz, Austria
kalkofen@icg.tugraz.at

Andreas Hein
University of Oldenburg
Oldenburg, Germany
andreas.hein@uol.de



a) Input video

b) Automatic acquisition of the 2D trajectory and 3D posture of up to two hands

c) Manual authoring to reconstruct a 3D trajectory

d) Augmented Reality Tutorial

Figure 1: The author chooses the input video (a). TutAR extracts relevant hand motion from the video, estimates the hand posture and creates an animation (b). With an authoring tool, the author can reconstruct a 3D hand trajectory relative to the human body (c). The animation of the 3D hand will be displayed on the registered place and plays synchronously with the motion in the video on an OST-HMD (d).

ABSTRACT

With Augmented Reality (AR) on Optical See-Through Head-Mounted Displays (OST-HMD), users can observe the real world and computer graphics at the same time. In this work, we present the design and implementation of TutAR, a pipeline that semi-automatically creates AR tutorials of 2D RGB videos with hands-only procedures such as cardiopulmonary resuscitation (CPR). TutAR extracts relevant 3D hand motion from the input video. The derived hand motion will be displayed as an animated 3D hand relative to the human body and plays synchronously with the motion in the video on an OST-HMD.

VRCAI '18, December 2–3, 2018, Hachioji, Japan

ACM ISBN 978-1-4503-6087-6/18/12…$15.00
https://doi.org/10.1145/3284398.3284399

CCS CONCEPTS

• **Human-centered computing** → **Mixed / augmented reality**;
• **Computing methodologies** → *Reconstruction*;

KEYWORDS

Augmented Reality, Medical Education, Video Tutorials, Motion extraction

1 INTRODUCTION

Augmented Reality (AR) extends the real environment with virtual objects. An AR application has to consider virtual elements relative to real-world objects which makes creating AR applications often a complex task. Creating AR applications is hard. It requires mastering the technical components of an AR system. Especially creating life-like animations for hands that fit the user's real hands is often a time-consuming and challenging task. The author needs to have in-depth knowledge of 3D modeling tools while being able to master the technical components of an often complex AR system. Therefore, creating AR applications is often done by specially skilled people only.

Videos for many procedures are readily available and easy to produce. Especially, a large amount of videos tutorials exist in crowd-sourced databases providing a large amount of already authored movements. However, it can be challenging to follow instructions from video tutorials precisely. They must infer 3D movement, speed, and velocity only from 2D video cues. Also, the appearance of objects in the video may differ from the real world, making it difficult to identify matching landmarks. The user's view and the view from the video might be different which exacerbate the problem.

In this work, we focus on videos that convey hand motion. To make this information available to 3D AR, we introduce **TutAR**, a system which is capable of automatically transferring videos of hand procedures into 3D AR applications using minimal user input. As demonstrated in Figure 1, TutAR extracts a 3D hand model and its movements from 2D videos, and it displays the 3D information registered to real-world objects by using an OST-HMD.

2 RELATED WORK

Automatic approaches to generate AR applications go back to the pioneering work of Feiner et. al [3] from 1993. They describe Knowledge-based Augmented Reality for Maintenance Assistance (KARMA), a prototype of a scripting-based AR system for maintaining laser printer based on a see-through HMD.

A popular approach consists of overlaying video material on the user's view. Some works, including [4, 8], merely overlay a 2D RGB video of a reference performance directly onto the user's view on the AR display. Overlaying the original video has the advantage of showing unaltered content. This approach also does not require a lot of authoring. However, by just overlaying a 2D video, the user will not be able to have a free choice of the viewpoint limiting its practical value.

Another approach consists of generating AR applications out of two-dimensional technical documentation. Mohr et al. [5] describe a system to automatically compute a diagram's layout and create an animation of its explosion. It integrates the animation instructions into a real-world environment. However, these approaches only work with the straight movement for removing parts from a CAD model.

For AR applications which require more complex movement, motion capturing was used. Chi et al.[2] create instructions for gestural commands from complex motion captured data. The instructions are displayed in 2D for effective communication. However, their system does not display 3D instructions in AR.

Our approach is inspired by the work of Mohr et al. [6]. However, while their system creates AR tutorials showing tools with surface contact from 2D RGB videos, TutAR creates tutorials of hand movements from RGB video tutorials.

3 SYSTEM OVERVIEW

Figure 2 illustrates the pipeline of our proposed solution. Automated steps are colored blue, whereas steps which require user input are colored red. We are aiming to automate as many steps as possible. However, we rely on input from the user to register the AR tutorial to the real world object. While authoring, the user divides the video tutorial first into segments, each describing a single task. In each segment, TutAR acquires motion of up to two human hands. For

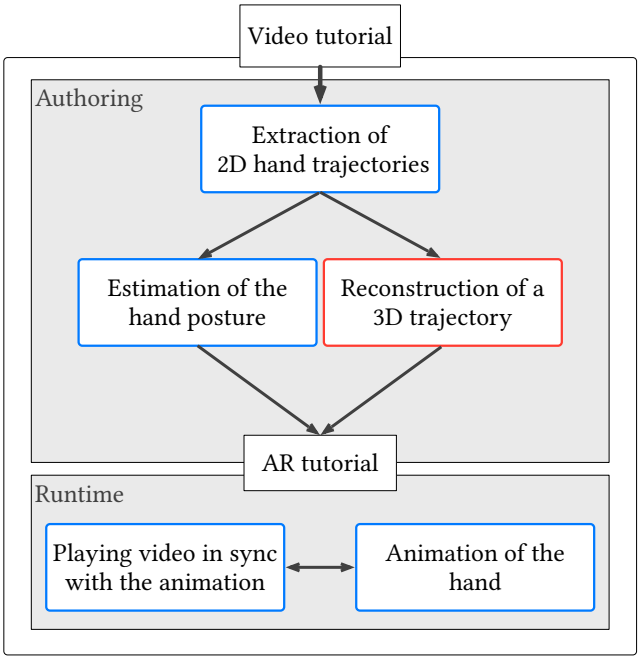


Figure 2: The pipeline of TutAR. Automated steps are colored blue, whereas steps which require author input are colored red.

acquiring hand motion, we are using OpenPose [1, 9, 10], a library for body keypoint detection allowing acquiring the position for 21 keypoints for each hand in image space. Subsequently, we use Hand3D [11] to retrieve the 3D hand posture. Finally, the user registers the extracted 3D hand motion to the real object (Figure 1 (c)). However, just extracting 2D hand motion in a video in image space is not enough to create a useful 3D AR application. Also, it is difficult to estimate the pose of an object in 3D space in a 2D image with an unknown environment. Thus, an authoring tool is needed.

Since we are focussing on tutorials for medical education which are using hands as the main tool on the human body, the authoring tool involves around creating a 3D trajectory relative to the human body. An authoring tool in which the author is asked to place the hand relative to the manikin is shown in Figure 1 c. The left part displays the left hand (green) and the right hand (red) on top of a manikin. On the right is a frame of the input video. After adjusting the hands of interest, the author can add the pose of the hands of the current frame as a keyframe. The author can add new keyframes when the position of the hand should be corrected. Between each keyframe, the pose will be interpolated.

Our prototype creates AR tutorials of 3D hand motion for the Microsoft HoloLens. We have built an example application based on a video of a medical training procedure. Our application renders the extracted 3D hands relative to a medical manikin to train a reanimation procedure.

4 EVALUATION

Our system is able to present the original video and the extracted 3D hand motion in sync. However, it is not clear which data to use

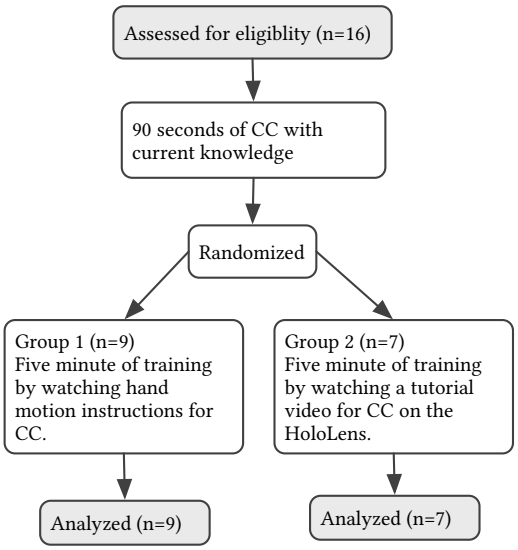


Figure 3: Flow chart of design and recruitment of participants.

at which point in time. Therefore, we have conducted an initial test to compare the learning effectiveness of watching a video vs. watching merely an AR animation. We also aimed to observe users in performing an AR guided task. We assessed the performance difference of cardiopulmonary resuscitation (CPR), particular the chest compression (CC) with the use of TutAR to a video.

Two crucial parameters of the CPR are the frequency at which the compressions are performed and the compression depth of the chest. The optimal CC has a frequency of 100 to 120 compressions per minute with a depth of 5 to 6 cm [7].

Figure 3 shows a flowchart of the procedure. Because of technical errors, the participants could have not been divided equally between both groups. The 16 participants were randomized into two groups: group 1 used the tutorial created by TutAR where they could only see the created AR animation. Group 2 saw a looped muted tutorial video for CC. The input video used for TutAR was the same as the video group 2 was watching. Both groups used the Microsoft HoloLens as an OST-HMD.

Overall, differences between group 1 and 2 were not significantly different for *compressions with correct rate* . While we could not prove that an AR animation is outperforming a video for *compressions with correct depth* either, we could make some interesting observations. Due to the small field of view (FoV) of the HoloLens, many participants noted that they could not see the augmented hands well without moving their head too far away from the chest. We expect that an OST-HMD with a bigger FoV would greatly improve the learning performance for this use case.

Also, some addressed a problem of depth perception because the augmented hands showed no impact on the chest of the manikin. We hypothesize that CPR does not benefit from a stereoscopic AR animation as a guide, because the user is looking at the augmentation from above which leads to a difficulty of perceiving the depth of the compression.

Further investigations are needed for evaluation and to see if benefits translate to an improvement in outcome, especially learning efficiency. We hypothesize that displaying both, the AR animation and the video in sync might be beneficial which will be conducted in a future user study.

5 CONCLUSIONS AND FUTURE WORK

TutAR is, according to our best knowledge, the first system for creating semi-automatically AR tutorials out of two-dimensional RGB videos for hand procedures. An example with CPR has been shown and examined in a user study. Our approach profoundly differentiates from existing work. We can reconstruct a 2D hand trajectory including the 3D hand posture. With additional author input, a 3D trajectory can be reconstructed.

In a future user study, we want to evaluate the learning performance of a revised prototype based on the lessons learned from our initial test. In the future, we want to extend the system to handle other types of video tutorials. For example, allowing to track the motion of whole human bodies or other types of tools.

6 ACKNOWLEDGEMENTS

This work is partly supported by the Austrian Science Fund grant No. P30694. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Computer Vision and Pattern Recognition*. IEEE, 7291–7299.
[2] Pei-Yu Chi, Daniel Vogel, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2016. Authoring Illustrations of Human Movements by Iterative Physical Demonstration. In *Annual Symposium on User Interface Software and Technology*. ACM, 809–820.
[3] Steven Feiner, Blair Macintyre, and Doree Seligmann. 1993. Knowledge-Based Augmented Reality. *Commun. ACM* (1993), 53–62.
[4] Michihiko Goto, Yuko Uematsu, Hideo Saito, Shuji Senda, and Akihiko Iketani. 2010. Task support system by displaying instructional video onto AR workspace. In *International Symposium on Mixed and Augmented Reality*. IEEE, 83–90.
[5] Peter Mohr, Bernhard Kerbl, Michael Donoser, Dieter Schmalstieg, and Denis Kalkofen. 2015. Retargeting Technical Documentation to Augmented Reality. In *Conference on Human Factors in Computing Systems*. ACM, 3337–3346.
[6] Peter Mohr, David Mandl, Markus Tatzgern, Eduardo E Veas, Dieter Schmalstieg, and Denis Kalkofen. 2017. Retargeting video tutorials showing tools with surface contact to augmented reality. *Conference on Human Factors in Computing Systems* (2017), 6547–6558.
[7] Gavin D. Perkins, Anthony J. Handley, Rudolph W. Koster, Maaret Castrén, Michael A. Smyth, Theresa Olasveengen, Koenraad G. Monsieurs, Violetta Raffay, Jan Thorsten Gräsner, Volker Wenzel, Giuseppe Ristagno, Jasmeet Soar, Leo L. Bossaert, Antonio Caballero, Pascal Cassan, Cristina Granja, Claudio Sandroni, David A. Zideman, Jerry P. Nolan, Ian Maconochie, and Robert Greif. 2015. European Resuscitation Council Guidelines for Resuscitation. *Resuscitation* 95 (2015), 81–99.
[8] Nils Petersen and Didier Stricker. 2012. Learning task structure from video examples for workflow tracking and authoring. *International Symposium on Mixed and Augmented Reality* (2012), 237–246.
[9] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *Computer Vision and Pattern Recognition*. IEEE, 1145–1153.
[10] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *Computer Vision and Pattern Recognition*. 4724–4732.
[11] Christian Zimmermann and Thomas Brox. 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. In *Computer Vision and Pattern Recognition*. IEEE, 4913–4921.