

# Webcam-based Hand- and Object-Tracking for a Desktop Workspace in Virtual Reality

Sebastian Pape  
sebastian.pape@rwth-aachen.de  
RWTH Aachen University  
Aachen, Germany

Torsten W. Kuhlen  
kuhlen@vr.rwth-aachen.de  
RWTH Aachen University  
Aachen, Germany

Jonathan Heinrich Beierle  
jonathan.beierle@rwth-aachen.de  
RWTH Aachen University  
Aachen, Germany

Tim Weissker  
me@tim-weissker.de  
RWTH Aachen University  
Aachen, Germany

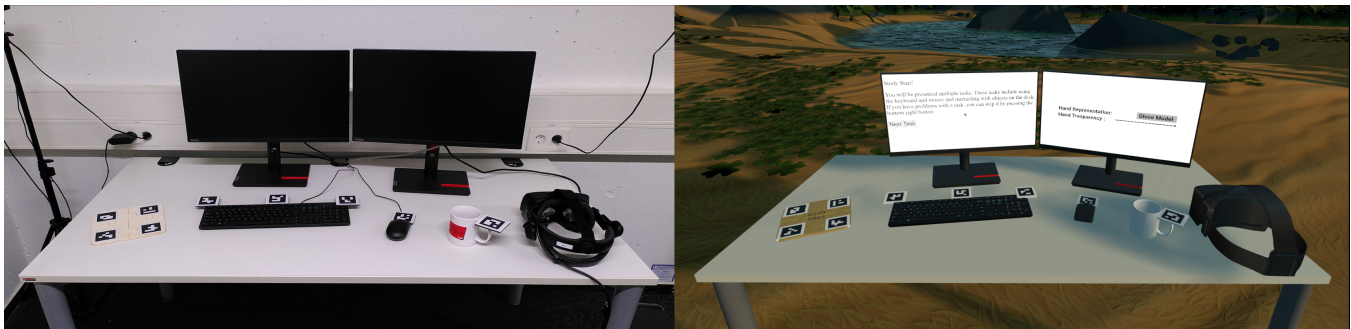


Figure 1: Overview of the physical (left) and virtual (right) setup in comparison.

## Abstract

As virtual reality overlays the user's view, challenges arise when interaction with their physical surroundings is still needed. In a seated workspace environment interaction with the physical surroundings can be essential to enable productive working. Interaction with e.g. physical mouse and keyboard can be difficult when no visual reference is given to where they are placed. This demo shows a combination of computer vision-based marker detection with machine-learning-based hand detection to bring users' hands and arbitrary objects into VR.

## CCS Concepts

• **Computing methodologies** → **Tracking; Matching;** • **Human-centered computing** → **Virtual reality; Keyboards.**

## Keywords

Virtual Reality, Hand-Tracking, Object-Tracking, Physical Props, Webcam

## ACM Reference Format:

Sebastian Pape, Jonathan Heinrich Beierle, Torsten W. Kuhlen, and Tim Weissker. 2024. Webcam-based Hand- and Object-Tracking for a Desktop Workspace in Virtual Reality. In *ACM Symposium on Spatial User Interaction (SUI '24)*, October 7–8, 2024, Trier, Germany. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3677386.3688879>

## 1 Introduction

For general office tasks virtual- and mixed-reality technologies can be used to create a more focused working environment that removes distractions caused by co-workers or other influences, leading to more productive work. Additionally, virtual screenspace or even 3D visualizations can be placed directly in front of the user to facilitate more direct interaction. Furthermore, virtual offices can increase social engagement with co-workers in cases of prolonged remote work.

A big challenge in all of these scenarios lies in the coupling of real-world objects and the virtual workspace, as every object in the real world needs a synchronized virtual representation to be visible to the user in the virtual world.

In our demo, we present a system that is based on two readily available webcams detecting the user's hands in addition to fiducial markers attached to objects on the desk.

## 2 System Setup

The system is based on two, not necessarily similar, webcams that capture stereo images of the workspace. In our demo, the workspace will be represented by a desktop workspace, as shown in Figure 1.

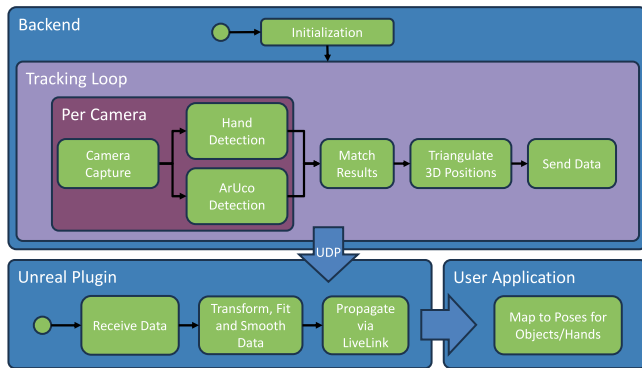
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SUI '24, October 7–8, 2024, Trier, Germany

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1088-9/24/10

<https://doi.org/10.1145/3677386.3688879>



**Figure 2: Schematic overview of the system, showing the different steps and parallel processes.**

With the dual camera setup, objects on the table and the user’s hands will be tracked and synchronized with a virtual counterpart in the VR demo. While the presented system supports more than two cameras, the demo will be performed with a dual camera setup, as it represents the least amount of setup actual users would need to make. Currently there exist multiple hand-tracking solutions for usage in VR, some of which are built into the headsets themselves, but the usage of external cameras or even a combination of both can significantly reduce the problems of occlusion and the accuracy of the whole system. As the used framework is executed on RGB camera images the system is very resilient regarding lighting conditions, background materials and distance to the camera.

In Figure 2 an overview of the technical setup of the system is given. The single parts of this system are described in the following sections.

## 2.1 Backend

For synchronization, detection and distinction of objects, the system uses ArUco markers [2], which are printed onto foamboard and are attached to objects. As a first step in the backend, see Figure 2, one image per camera is used for calibration purposes. In these images the markers of all objects are detected, matched and used to determine the position and orientation of the cameras. Afterward, in a loop, each camera’s image is captured and processed in parallel for hand and marker detection. For the hand-detection Google’s MediaPipe framework [1] is used to detect 21 *landmarks* on each hand and decide which hand is a right or left hand. After both detection steps, the resulting image points are matched and triangulated to determine their position in space. As a last step, the results are serialized and passed to the Unreal Engine via UDP for further processing.

## 2.2 Frontend: Unreal Plugin + User Application

In the Unreal Engine, the data is deserialized and transformed into the Unreal coordinate system and then further propagated via the LiveLink system, a unifying system for tracking data in the Unreal Engine. In case multiple markers are attached to a single physical object, these are then fit using a least-squares method to find the

most fitting representation. This leads to more a robust detection in case of partial occlusion or bad lighting conditions.

Based on the data in the LiveLink system the user is free to animate hand models or virtual objects in their application.

## 3 Initial User Evaluation

The system was initially evaluated in an expert review with  $n = 8$  participants. For this, the experts tested the system prototype in a semi-structured interview while executing the tasks that are presented in this demo (see section 4). The experts voiced generally positive feedback while acknowledging its weaknesses in terms of occlusion tolerance and stability. In addition to the semi-structured interviews the NASA-TLX [3] and the User Experience Questionnaire (UEQ) [4] were given to the participants. The NASA TLX resulted in a raw average score of 32.9 (SD 11.9). Additionally, participants rated the prototype with an Attractiveness of 1.08, Perspicuity of 1.47, Efficiency of 0.875, Dependability of 0.66, Stimulation of 1.22, and Novelty of 1.16 on the scales of the UEQ. Based on these results, the system was deemed useful and future research could lead to further significant improvements.

## 4 Core Features of the Demo

Visitors can test and experience the prototype system while performing some of the tasks that were used in the expert review. These include the manipulation of physical props (coffee mug, jigsaw puzzle, mouse, keyboard) and interaction with virtual desktops (Multiple mini-games and a typing test) in sessions lasting 2-5min. Visitors will also be able to see the user’s view on one of the monitors while waiting. The demo should not only show the possibilities of readily available computer vision algorithms but also demonstrate the weaknesses of such systems, hopefully leading to discussions and feedback for future developments of such systems.

## 5 Conclusion and Future Work

In future iterations of this prototype, we would like to further increase the accuracy and reduce the latency of the tracking. Furthermore, the MediaPipe framework also offers a full body detection module, which could be used to further increase the sense of embodiment in applications, where it is needed. Additionally, we would like to extend the current prototypical system to fuse the hand-tracking of modern headsets with the output of the system to minimize occlusion problems and increase hand-tracking stability.

## References

- [1] Camillo Lugaresi et al. 2019. MediaPipe: A Framework for Perceiving and Processing Reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.
- [2] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* 47, 6 (2014), 2280–2292. <https://doi.org/10.1016/j.patcog.2014.01.005>
- [3] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- [4] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. *USAB 2008* 5298, 63–76. [https://doi.org/10.1007/978-3-540-89350-9\\_6](https://doi.org/10.1007/978-3-540-89350-9_6)