ProcessAR: An augmented reality-based tool to create in-situ procedural 2D/3D AR Instructions

Subramanian Chidambaram
schidamb@purdue.edu
Purdue University
IN, USA

Hank Huang
huang670@purdue.edu
Purdue University
IN, USA

Fengming He
he418@purdue.edu
Purdue University
IN, USA

Xun Qian
qian85@purdue.edu
Purdue University
IN, USA

Ana M Villanueva
villana@purdue.edu
Purdue University
IN, USA

Thomas S Redick
tredick@purdue.edu
Purdue University
IN, USA

Wolfgang Stuerzlinger
w.s@sfu.ca
Simon Fraser University
BC, Canada

Karthik Ramani
ramani@purdue.edu
Purdue University
IN, USA

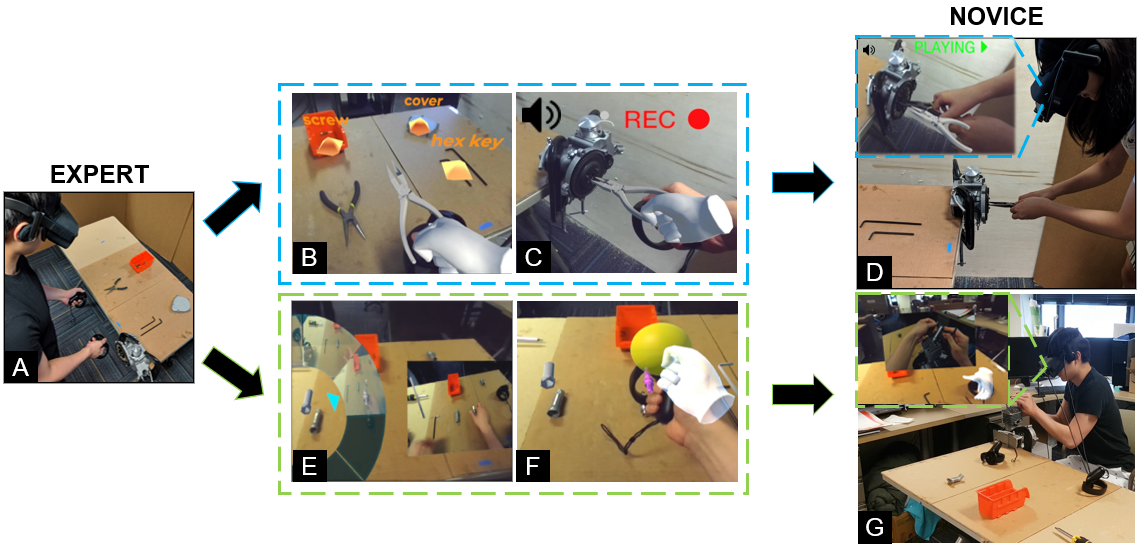

Figure 1: A Subject Matter Expert (SME) creating AR instructions in-situ within AR (Fig A). ProcessAR detects the objects in the real world with computer vision and spawns the corresponding virtual tool (Fig B). The SME moves the virtual tools in real space, and the motion of this virtual tool is recorded along with audio to create an AR instructional checkpoint (Fig C). Embedding captured 2D videos within the AR work environment (Fig E & F). The created AR content in 3D (Fig D) and 2D (Fig G) is transferred to the novice user, who then performs the task.

DIS '21, June 28-July 2, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8476-6/21/06…$15.00
https://doi.org/10.1145/3461778.3462126

ABSTRACT

Augmented reality (AR) is an efficient form of delivering spatial information and has great potential for training workers. However, AR is still not widely used for such scenarios due to the technical skills and expertise required to create interactive AR content. We developed ProcessAR, an AR-based system to develop 2D/3D content that captures subject matter expert's (SMEs) environment-object interactions in situ. The design space for ProcessAR was identified from formative interviews with AR programming experts and SMEs, alongside a comparative design study with SMEs and

novice users. To enable smooth workflows, ProcessAR locates and identifies different tools/objects through computer vision within the workspace when the author looks at them. We explored additional features such as embedding 2D videos with detected objects and user-adaptive triggers. A final user evaluation comparing ProcessAR and a baseline AR authoring environment showed that, according to our qualitative questionnaire, users preferred ProcessAR.

CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**.

KEYWORDS

Augmented Reality; Authoring ; tutorials; Computer Vision

ACM Reference Format:
Subramanian Chidambaram, Hank Huang, Fengming He, Xun Qian, Ana M Villanueva, Thomas S Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2021. ProcessAR: An augmented reality-based tool to create in-situ procedural 2D/3D AR Instructions. In *Designing Interactive Systems Conference 2021 (DIS '21), June 28-July 2, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3461778.3462126

1 INTRODUCTION

Spatial ability has been defined as "the ability to generate, retain, retrieve, and transform well-structured visual images in spatial relations among objects or in space" [36, 68]. Such spatial abilities play a critical role in our everyday lives and also at work, in tasks such as assembly, tool-manipulation, and navigation. Augmented reality (AR) is a technology that superimposes computer-generated virtual models in real-time on a user's view of the real world. AR has been shown to be a reliable mode of instructional training, improving speed and reliability, minimizing errors, and reducing the cognitive load of the user [3, 37, 41], especially for spatially distributed instructions.

Many tasks in an industrial or manufacturing environment are spatial in nature. Critically, in many of these industrial sectors, companies are unable to keep up with getting new employees trained as fast as experienced people are retiring [13]. Often called the "*skills-gap*", the resulting problem has reduced the supply of well-trained workers for the manufacturing industry, which is widely recognized as a critical element to be addressed by future work-related research [29]. Currently there are three widely popular modalities of training employees in spatial tasks: one-on-one, paper/sketch-based, and video-based instruction, with AR being explored as a potential fourth alternative. The current approaches to instruction creation require capturing user actions through written instructions, sketches, pictures, and videos from the real world and require extensive editing to produce usable instructions. Only a few large corporations such as Boeing [5] and Lockheed [38] have been able to afford the widespread use of AR to train their workforce. One of the primary reasons for the lack of widespread use of AR can be attributed to the complexity and technical knowledge required to develop AR instructional systems. For work instructions, current authoring tools require AR programming experts working in tandem with subject matter experts (SMEs) to capture the relevant knowledge and skills.

AR applications often have to consider the location of the virtual objects relative to the variable environment, increasing the complexity of creating AR content [16]. In addition, knowledge of 3D modelling and animation are required if usable instruction sets are to be created by the author [41]. This is a tedious, time-consuming, expertise-intensive activity. It involves multi-person collaboration, which is not only costly but also a major impediment to AR content creation and widespread success of AR itself. However, by utilizing the unique position that AR occupies on the spectrum from virtual to reality, i.e., right in the middle between the two extremes, our work addresses a major part of the challenge of AR instruction creation, towards eventually removing such impediments.

We present ProcessAR, an AR-based system to develop 2D and 3D procedural instruction for asynchronous AR consumption by capturing user/expert actions. In this work our goal is to empower SMEs to directly author AR instructions, by reducing the complexity and easing the load required for creating AR content. We achieve this by using principles similar to programming by demonstration (PbD) [33]. Instead of performing the actions in a entirely virtual or entirely real environment, the expert authors demonstrate the actions by manipulating virtual tools or parts–identified by computer vision–overlaid onto the physical world.

Currently, many industries still use one-on-one training to train their employees for tasks such as machine operation and manufacturing. Although reliable, this mode of training is inefficient in terms of time, cost, and scalability. One-on-one training requires active feedback and communication between SMEs and novices. Past work such as Loki [67], Oda et al. [46], and Elvezio et al. [17] explored AR/VR based interfaces for remote one-on-one training in this area. However, unlike the synchronous nature of such instructions and training, ProcessAR explores the asynchronous nature of instructions, which relies on the ability to record and replay instructions at any time without the constant presence of an SME. Thus, ProcessAR requires a different set of interaction modalities relative to [46, 67].

By merging recent advancement in computer vision/AI algorithms such as You Only Look Once (YOLO) v1 [53], v2 [54], and v3 [55], the system does not need to know the a-priori position of virtual objects in the physical space, enabling the system to better and directly understand the current state of the environment. The user just looks at all the tools once in the beginning of the workflow, which enables the system to automatically identify the corresponding virtual tools to be overlaid. The virtual overlay enables the system then to automatically match the corresponding virtual objects with the real ones, eliminating the need for significant preprocessing, as required by previous work, such as AREDA [4], Fiorentino et al. [18], or Ong et al. [47]. Together with the Oculus Rift hardware, a ZED Mini [63] depth camera is used in our system to robustly determine the location of the tools/objects and eliminates the need for visual markers, unlike previous work such as Oda et al. [46].

ProcessAR also eliminates the need for transitioning between different modalities and interfaces to create and edit a video or a paper instruction. Working with SME spatial demonstrations and actions using virtual tools, ProcessAR directly enables authors to create AR content within an AR environment. Our aim with this work is to empower SMEs to create their own instructions through

natural interactions within the environment and without the large time requirements. In this way, we are paving the way towards making AR content creation more accessible and intuitive for the community.

The main contributions of our work are:

- Developed a workflow which enables SMEs to asynchronously create AR content (3D and 2D) and simply through demonstration. More importantly, our workflow enables retention of tool-manipulation trajectories, in the form of positional and rotation information, which are key to successfully train new workers in complex spatio-temporal tasks.
- Developed an AR-based authoring platform which empowers SMEs to become content creators without technical AR authoring know-how.
- Comparative user evaluation of the system, with two set of studies providing insights on the usability of our system to create AR instructions, in comparison to traditional asynchronous instructional media.

2 RELATED WORK

2.1 AR for Assembly Instruction

For nearly two decades, mechanical assembly operations have been of particular interest to AR researchers [52, 59]. Much previous work [4, 20, 24, 26, 47, 65] has proposed different approaches to solve the problem of authoring assembly instructions. For simple assembly operations such as Duplo block [24] or model vise [4] assembly, visually portraying the initial and final state of an object before and after assembly operation might be enough. But for more complex tasks, including tool manipulation, where the same tool could be used in different orientations (such as an Allen key), or illustrating the speed or posture require to perform different operations, merely stating the initial and final state for tasks is insufficient. For successful assembly instructions of such complex tasks, the instruction set must deliver information of tool/object movement and orientation, both with respect to time within the work space. We call this "spatio-temporal instructions". ProcessAR supports embodied authoring, which leverages the spatial and temporal nature of object manipulations and user interactions with them, while simultaneously providing the ability to view the instruction from multiple perspectives (3D authoring). This is a key contribution of our work, because previous work has failed to capture position and orientation of tools, thus disregarding crucial information for training workers in complex spatial tasks. In this context, ProcessAR plays a crucial role in showing "how" to use tools, which has not being addressed by prior work [19, 39, 62, 69]. This is done by enabling the author to demonstrate tool interaction with the aid of a virtual tool, and captures such interaction for replay at a later time.

2.2 3D authoring

Although not designed for asynchronous learning environments, previous projects [46, 67] are capable of representing spatio-temporal 3D AR animation. However, these tools leverage only remote training and synchronous learning instead of an asynchronous modality–which requires a live instructor and therefore does not scale to many users. While many features, such as using virtual replicas for demonstration [46], embedding 2D content [39, 49], and point-of-view indication [64] have been explored and studied independently in the past, such as in Cao et al. [8], their synthesis has not been validated nor had insights been drawn from an integrated 2D & 3D AR in-situ authoring platform such as ProcessAR. Thus a thorough study of such an integrated system–as in ProcessAR–is necessary for this approach to be validated and to provide value to the HCI community. ProcessAR enables an integrated 3D/2D AR authoring environment, which presents a unique set of previously unexplored opportunities and challenges for tool interaction and manipulation.

Ong et al. [47] and Radkowski [51] present workflows that allow the user to use their bare hands to interact with virtual objects. While these preserve the natural and intuitive interactions our work desires to mimic, they rely on artificial markers placed around the hands. These markers are susceptible to occlusion and subsequent loss of tracking, preventing a robust implementation required for spatio-temporal tasks, especially in the workplace. Several examples such as AREDA [4], [46, 56] and Ha et al. [25] use line-of-sight markers, which can disrupt the task to be demostrated. These approaches require significant pre-processing and are susceptible to occlusion by hands and other objects caused by limitations of computer vision systems [34], which is a major problem for the capture of smooth user interaction. ProcessAR is novel in that it preserves natural interaction by allowing the SME to create content through demonstration and bypasses occlusion issues by matching the physical objects with their virtual counterparts at all times.

2.3 AR through Motion Capture

Teach me how! [20] presents a PbD based AR authoring system that utilizes a Kinect depth sensor and a projector to understand and augment assembly for (only) 2D AR instructions. DemoDraw [11] is a system that translates speech and 3D joint motion into a series of key pose demonstrations and illustrations. YouMove [1] allows users to record and learn physical movement sequences, primarily for dance. Physio@Home [66] is a system that guides people through pre-recorded physiotherapy exercises using real time visual guides and multi-camera views. Another approach [32] is limited to capturing body motions. ProtoAR [43] enables users to create AR content with rapid physical prototyping using paper and Play-Doh. ProtoAR provides a new mobile cross-device multi-layer authoring and interactive capture tools to generate mobile screens and AR overlays from paper sketches, and quasi-3D content from 360-degree captures of clay models. All these approaches are limited to capturing and rendering body motion tasks, and are not concerned with hand-held manipulation of tools, which is essential for training in an industrial environment.

2.4 AR through 2D Video

Several other past systems in this space, such as PTC's Vuforia Capture [49] or [9], only capture 2D video, to be embedded later over the real world, and require significant post-processing, which can often be limiting and lose context, as described above. Similarly, and unlike ProcessAR, other approaches [10, 31] are limited to 2D video for instructional delivery and do not involve AR. The idea of obtaining AR instructions from video have been explored previously in Mohr et al. [41] and TutAR [16]. Such work is limited

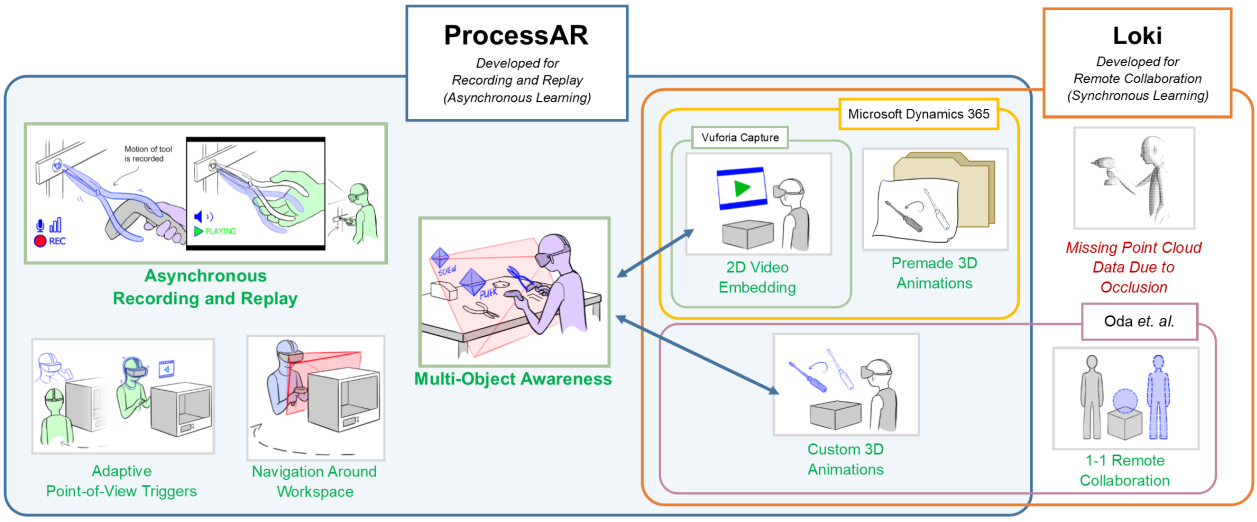

Figure 2: Comparison of ProcessAR with respect to different dimensions of the design space occupied by various existing AR authoring system. This figure visualizes the various features of each system, highlighting both elements that are shared and unique to each system.

to generating AR for applications only to situations where a source video is available. Also, Mohr et al. [41] deals only with surface contact, while TutAR [16] requires manual user annotation during post-processing of the creation process.

Instead of providing 3D instructions, such as Goto et al. [23], Petersen et al. [48], multiple systems directly embed 2D videos into the environment. "YouDo, I-learn" [14] records object usage and replays the recorded video upon a gaze-based trigger. Due to the 2D nature of the videos, the separation between the screen and the location of the task leads to a higher mental effort for hand eye coordination [57].

With purely video-based knowledge sharing, users often miss the spatial nature of task instructions and can lack temporal context. To supply the missing information, video instructions often must be captured from multiple perspectives, then edited together for effective instruction. This requires time and expertise in video editing and transforming content between software tools. Ego-centric videos suffer from the same limitation. In contrast, well-authored 3D animated AR instructions created through ProcessAR are able to holistically and directly capture 3D tool operation, without requiring multiple perspectives. This reduces development time while maintaining quality of instructions.

2.5 AR through 2D Interfaces

Mohr et al. [40] explored the idea of extracting AR instructions from paper-based technical manuals but was only capable of generating tutorials from instructions with straight motions and where the source was available. Quick and easy editing or modifying instructions is not possible with their approach. Recent work Saquib et al. [57] presented an AR authoring environment to trigger a 2D augmented video performance. This work is currently only capable of delivering 2D AR animations which have been pre-programmed by the user in a 2D interface.

Unlike previous work such as [27, 39, 42, 49] which rely on post-processing through a PC app, ProcessAR does not rely on external processing tools. The elimination of additional pre- and post-processing steps supports faster content authoring. Compared to traditional methods, creating AR content within the ProcessAR platform enables expert users in a manufacturing environment to create 3D AR content quickly, more naturally, and intuitively.

3 DESIGN SPACE EXPLORATION

After exploring the past literature in this space from a research perspective, we interviewed 12 expert manufactures, educators, and AR programmers to identify the real world constraints that they currently face for large scale deployment of AR.

3.1 Formative Interview

The first interviewees were six subject experts in the area of manufacturing working for global firms, specializing in the manufacturing of trucks, electronics, telecommunication equipment, pharmaceutical machines, automobiles and engines. Additionally, we interviewed three educators in the area of manufacturing and three AR programmers. They shared their past experiences, current constraints and industry practices they face in their industry regarding training new employees, and what they expect from a product. We distilled and present the most relevant information here, which along with a preliminary design study to understand the users, were used to set a group of design goals.

The experts defined that *Spatial Tasks* in manufacturing environments usually are a series of diverse human-tooling and human-machine interaction tasks happening at various locations in a large spatial environment. They argued that although multiple tutoring modalities have been adopted in human worker training, it remains challenging to author tutorials for Spatial Tasks. Currently one of

the most reliable ways to instruct someone is via one-on-one learning. However, they feel it severely limits their ability to grow or replace their quickly aging workforce and deal with high turn-over rates. The cost of scaling up one-on-one training to the required levels is much higher than most companies can afford. Hence, it is key to eliminate the reliance on the constant presence of the expert (i.e., a synchronous process) during training. Thus, an asynchronous system became an important part of our goals.

Other comments made by the experts which also supported the design of ProcessAR were:

- Another challenge for Spatial Task is the large space that has to be navigated between unit sub-tasks. That is, a user has to move within the work space.
- During training the experts emphasize a clean work environment, without clutter. As safety is an overarching requirement the employer would like to encourage this as a core value.
- A typical Spatial Task includes unit sub-tasks that hold significantly different attributes resulting in different requirements from a human worker. Thus a system that procedurally breaks down a task into sub tasks with vastly different attributes was explored.

3.2 Preliminary design study

To better understand the users interaction and strengths and weakness of different instructional media, we asked Expert users to create instructions for three different Spatial Tasks. We asked the users to create instructions as (1) paper/sketch based instructions, (2) videos, and (3) AR instructions. Initially we did consider one-on-one instructions; however, based on the formative interviews, where experts identified one-on-one to be highly inefficient in terms of time cost, we decided to exclude one-on-one instruction and to perform the study with the three other modes of instructional transfer to arrive at a fair comparison.

For this comparison we built a preliminary AR authoring system (from here on referred to as "baseline") and compared it to the others options in terms of capability, usability, and ease-of-use with two lab-based studies. This baseline AR authoring system contained features described in section 4.4 and 4.5.1 of this paper. All other features where added based on design goals (section 3.4) obtained as a result of the observation from this preliminary design study and the formative interview described in section 3.1.

The first study was conducted with ten expert users (EUs) (all male) (see table 1), with expertise in spatial tasks such as Engine assembly, Bike repair, and Shelf installation to create procedural instructions. These tasks were chosen due to the spatial and procedural nature of the tasks. The engine assembly experts were recruited from the Formula SAE [30] team. Two of the bike experts were from a local bike club and the other expert had experience working at a bike repair store. The experts used for shelf installation were woodworking hobbyists. To validate the usability of instructions created, a second study was conducted on 16 (4 female, 12 male) novice users.

User Study 1: Due to the embodied nature of AR instruction creation, we hypothesized that the time to create AR content would be lower compared to the other two modes, even if the EUs did not have AR experience prior to the study. We also believe that the system would lower the mental effort while authoring. The task for the experts was to assemble a 80cc gasoline motorbike engine (Refer fig.3 A), to remove and reinstall the front wheel of a bicycle (Refer fig.3 B), and to install a shelf to a well-secured mock wall (Refer fig.3 C), respectively. All EUs were paid a sum of US $60.

Figure 3: Expert Users creating procedural instructions for assembling an engine (A), repairing a bike (B), and installing a shelf (C).

The EUs were first briefed about the study and asked to fill a pre-survey questionnaire. All EUs were familiar with tools such as Microsoft Word and PowerPoint. Two EUs reported having created video instructions before but not for this particular task. Three EUs reported having used VR HMDs for playing games. During the study the EUs were asked to create instructions in all three modes of instructional transfer (instruction mode was manipulated within-subjects). The study was split into three sessions, each of which lasted for 2 hours, conducted on different days to prevent fatigue.

Session 1: The EUs were asked to create paper-based instructions for the tasks. They were provided access to sketching tools, such as Autodesk Sketchbook [60] or Inkscape [28], and a Microsoft Surface Pro 6. They were also provided with a Pixel 3, for capturing images, a bogen 3036 tripod with a phone mount, a computer with common document editing tools such as Microsoft Word, PowerPoint, and GIMP [21]. All users were given the choice to receive 25 minutes of training on any of the above mentioned tools. None of the users utilized the optional training period citing familiarity with the tools of their choice. Hence session 1 was conducted without any training. All EUs worked alone with the help of the tripod for a different point of view, depending on the situation.

Session 2: The EUs were asked to create an instruction video. They were given access to the video editor Camtasia 9 [7]. All EUs underwent a 25 minute tutorial on the basic features of the video editor. As part of the training, all the EUs were asked to edit a sample to confirm their learning of the tool. A microphone, the Pixel 3, and tripod from session 1 were again provided to the users for video capture to create the instructions. Completion time (CT) was recorded for each user while they recorded the video, transferred the videos to a computer, and finally during editing. EU8 and EU9 recorded voice simultaneously and the rest of the users recorded their voice post-hoc and added them to the video later.

Session 3: The EUs were asked to create AR instructions. All users were given a 25 minute training prior to the task. The users were first exposed to Oculus Touch Basics [44]. Then, in order to train them to become familiar with interacting with virtual tools, such as grabbing and re-orienting, the users played a VR app called Wrench: Engine Building Demo [44] for 5 minutes. Finally, the users

Table 1: Expert User's background information along with SUS Scores for the AR mode

User Id	Years of experience/Backgrounds	Expertise	SUS
1	4-FSAE & Engine	Engine	82.5
2	7-FSAE & Engine	Engine	85
3	6-FSAE & Engine	Engine	70
4	5-FSAE & Engine	Engine	80
5	15-Assembly	Engine	70
6	8-Bike Club	Bike	67.5
7	8-Bike Club	Bike	82.5
8	3-Bike Club	Bike	67.5
9	11-Hobbyist	Shelf	52.5
10	9-Hobbyist	Shelf	75

Average - 73.25

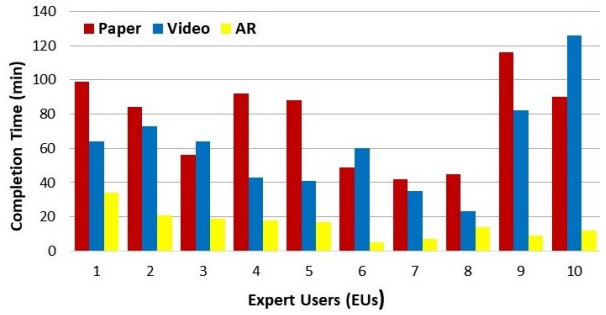

Figure 4: Authoring time to create various modes of instructions.

were exposed to ProcessAR for 15 mins as part of the training. For this, a one-on-one demo was provided on how to interact with the system. The trained users were then asked to create AR instructions using ProcessAR for the specific task of their expertise. Each of their demonstrations were then saved in the form of checkpoints (explained in section "Virtual Object Recording").

Post Survey: After the three sessions were complete, the users were asked to complete a questionnaire containing a system usability survey [6] and a few questions on their preferences among the three modalities. They were also asked to rank-order the three modalities in terms of preference. Additionally, every user was asked to provide the reasoning behind their rank ordering and preferences. An equal number of uses rated using AR and video as their first preference, with Paper instructions being rated as the least preferred.

User Study 2: The purpose of study 2 was to validate the usability of the instructions what were created in Study 1. To validate the usability of the created instructions, we invited 16 Novice Users (NUs) to be part of a second study. The objective for the NUs was to follow the instruction set given to them and to complete the task. Every NU performed all three tasks but utilized three different modes of instructions to complete the tasks. To prevent the NUs from learning aspects of the tasks, we used the three different modes of instructions for the three different tasks. We monitored the NUs and recorded successes/failures of each task and sub-task.

3.3 Observations and results

The results from the user studies provide evidence for the benefits that in-situ AR authoring can provide. The timing data for content creation provides clear evidence that, in all cases, the amount of time that the experts required for authoring instructions was reduced compared to the other alternatives (refer fig. 4). The ratio of reduction depends on the complexity of the task. The novice success percentages from study 2 corroborate the fact that all instructions created by the experts are indeed usable. The data did not reveal any significant difference in the ability to complete the task among the three modalities. All modes have a success rate higher than 85% (mean= 96.55; SD= 5.57), see the supplementary material. This high success rate is seen as evidence that all generated instructions can be considered to be usable. Additional observations of the NUs for sub-tasks specific to tool manipulation, such as the use of a hex key and combination wrenches, demonstrated that no mistakes were made when using the AR instructions. One of the NUs made the following comment: "I think AR was very cool. It even showed me where to put that thing [the hex key] and even how to turn it". Comparatively, there were at least two recorded cases of errors in tool usage with the paper instructions. We see this as evidence that AR instructions created within AR (an in-situ system) can directly have an impact in terms of error reduction. This can be corroborated if more complicated spatial tasks with a larger user group were to be conducted.

Overall we received positive feedback for ProcessAR. The average SUS score for ProcessAR was 73.25 (table 1), which implies that the system is definitely usable, but that there is room for improvement. One of the EUs commented during the post-study survey regarding his preference for video over AR: "I would [have] liked to have used a video at the end to assemble the brake wire back into the housing. It is very precise and I wish I could have just capture[d] while I was doing it instead of showing [it] using the virtual model. I would have been able to feel it, instead of being shaky". Although it is more work intensive and tedious to create videos, they are capable of capturing the finer details, also in motions. Thus we believe that combining the strengths of an in-situ motion capturing AR systems, as our baseline system, and the ability of embedding videos could be a powerful combination, encouraging us to think in terms of adding multi-media AR support.

Also, we observed that some NUs were not looking at the region of interest where the AR instructions were rendered, at least at the beginning of a check point. This led us to explore two additional features that could guide the user to a point of view of interest and integrate such a feature into our current baseline system, namely adaptive point of view trigger and navigation around the work space (Section 4.5.4).

Based on the outcomes of our work, coupled with the outcomes from the formative interviews, and results from the literature, we set the following design goals and started development of ProcessAR with the preliminary AR system as a starting point.

3.4 Design Goals

- **D1. Spatial Movement and Spatial Awareness** The importance of spatial movement and awareness is evident from the results of our preliminary design study, the literature,

and expert interviews. We define spatial movement as the ability to understand how to interact with tools and objects within the work space, as well as how these tools and objects should interact with each other for successful completion of the task. For example, during an assembly task such as an engine block assembly, the sequence of parts to be assembled have to be strictly adhered to. Failing to follow this sequence will lead to failure of task completion. It is also important to address how the user themselves navigate within the work space. As the instructional medium of AR is rendered close to the area of interest, the instructions are rendered at different location within the work space. It is important that the system is aware of the next target location of instructional delivery and indicates it to the user.

- **D2. Multimedia Support** From the comments made by the EUs, we realized the importance of supporting a 2D-based medium, such as embedding 2D videos within the AR space to capture the advantages of both AR and Video. Thus, a system capable of supporting multiple media was explored. The importance of multimedia instruction is often disregarded in favour of focusing only on one type of medium, thus wasting the opportunity of providing multiple valuable options to content creators. For example, multimedia instructions would also enable SMEs to create 2D instructions when clear 3D animations cannot easily be recorded (e.g., in the case of animating flexible materials–wires and cables).
- **D3. Perspective Awareness** As observed in the preliminary study, NUs had trouble identifying the location of the next instruction within the workspace (e.g., if an instruction first takes place at the 12 o'clock position, while the next step has to take place at the 8 o'clock position). This led to users rewinding the instructions multiple times before they were able to catch up with the spatially distributed nature of the tasks. We address this problem in the next iteration of the baseline system (i.e., in ProcessAR).
- **D4. Asynchronous System** From the formative expert interviews and the literature, one of the major limitations of the current reliable instruction transfer system (i.e., one-on-one instruction) was identified to be the synchronicity of this process. As in the aforementioned case, a constant presence of the SME is required for a successful skill-transfer. Although asynchronous media, such as video and paper instructions, already exist, the immersive nature of spatial tasks is limited with these media and lacks spatial awareness. As such, AR is a useful medium to deliver such information, due to its immersion and its ability for object tracking and superposition of content. Synchronous AR systems for learning spatial tasks in remote interactions have be been explored (e.g., by Loki [67] and Oda et al. [46]) within controlled setups. Generalizing such approaches, we fill the research gap of asynchronous AR instruction with ProcessAR. The greatest difficulty with synchronous systems are the cost and scalability in training a larger group of novice users. Both of these problems are avoided with an asynchronous system.

4 PROCESSAR OVERVIEW

ProcessAR is an AR authoring system designed to improve the tutorial creation process of spatio-temporal tasks through leveraging the advantages of the combination of a system that recognizes important parts of the environment together with virtual object rendering. The system is operated first by an expert author that composes the tutorial and then later by a novice user that consumes it. We implemented a prototype to support this process which incorporates three phases: 1) visually scanning the surrounding physical tools once to locate them and later render the corresponding virtual models, 2) recording/editing expert motions with the virtual tools for the procedural task together with their vocal instructions, and 3) a novice consuming the tutorial in the form of 3D animations or embedded 2D videos and the expert's voice instructions.

4.1 Architecture

ProcessAR was deployed using an Oculus Rift HMD with a ZED Mini camera for AR pass-through attached, yet we also point out that ProcessAR is also capable of running on more modern headsets such as the Rift S and the Quest with a PC VR-compatible Oculus link. The AR system is powered by a gaming desktop computer with a Nvidia GeForce GTX 2080 Ti GPU and an Intel i7-9700K CPU. Our tracking setup consists of three Oculus tracking sensors in the environment as depicted in Figure 4 (Left). To minimize headset and controller occlusion in AR, two sensors were mounted on tripods, 2m above the floor, facing down and one sensor on the floor facing up. We developed the system with Unity3D, which runs AR video passthrough and object detection in real time. The egocentric AR view rendering is implemented via the ZED-Unity plugin [63], where camera frames and depth data are relayed from the ZED-Mini to a Unity3D program and displayed through the HMD at 60 frames per second (FPS). We used the OpenCV for Unity plugin [61] to deploy YOLO and perform computer vision computations all within Unity3D.

At runtime, a copy of each camera frame is pre-processed and fed to YOLO for real time object detection in Unity. Once an object is detected, its 2D vector positions are combined with depth-data from the ZED Mini to compute the corresponding 3D position, similar to the concept of a pinhole camera, which is then used to overlay virtual objects on top of the actual ones.

To allow interaction with virtual replicas of detected objects, we used the virtual hand representations that were bound to the Oculus controllers via the Unity Oculus SDK [45]. Using Unity, we also deployed a virtual control panel to allow users to monitor and edit their spatial task authoring process as well as the activation of object detection. A virtual laser pointer served as the main means to interact with the panel by tapping on a controller button.

4.2 Object Recognition

Real-time object recognition is essential to create responsive interactions between the user and the virtual replicas of objects in the environment. For this reason, we chose YOLO v3 [55] for object detection, also due to its speed and robustness. To perform object detection, we created our own image data set to train the model [1], today a standard method for machine learning.

[1]Please refer to the Supplemental Material

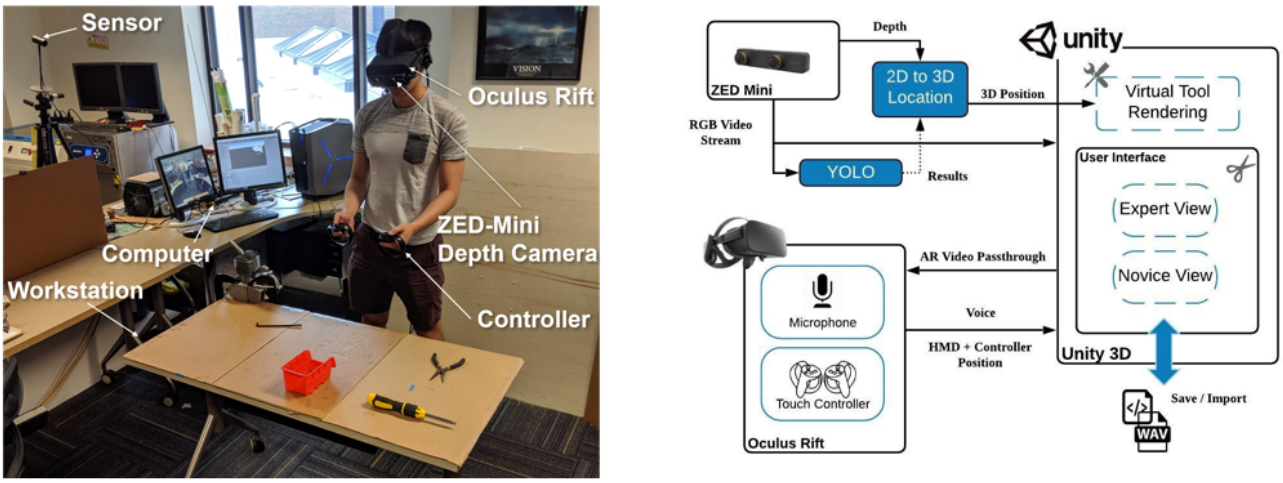

Figure 5: (left) Hardware setup for implementation. The modified Oculus Rift VR head-mounted display supports video pass-through with a ZED mini depth camera to work in AR mode. Oculus touch controllers are tracked using external sensors. (right) System Architecture: The workspace is observed by the ZED mini depth camera, from which a 2D video feed is sent into Unity 3D to render an AR environment.

To avoid the labour-intensive task of training objects, we relied on past research in this space. For each object, we collected approximately 2000 images using the method mentioned in LabelAR [35]. Specifically, we placed virtual 3D bounding boxes around the objects in AR and recorded the images from the ZED stereo camera. The 3D bounding boxes are projected onto images as 2D labels. The labeled data set were then trained, simplifying the workload.

During runtime, when the targeted objects come into the ZED Mini's FOV, ProcessAR recognizes the objects and provides detection information, which are the object's class and 2D bounding box. The results are processed in the background and the exact 3D location is generated based on a simple pinhole camera model. Once the 3D position is computed, ProcessAR renders a prefabricated computer-aided design (CAD) model corresponding to the detected tool object pulled from a repository within Unity for the user to manipulate.

Like YOLO, many other objection detection models visualize results by continuously drawing labeled bounding boxes. This may be suitable for video streams, but can be problematic in 3D space. Not only do bounding boxes lack spatial information, but it can be very computationally taxing to render a new virtual object based on the position detected by YOLO for every frame. To handle this issue, we implemented two methods: a virtual placeholder and training object substitution.

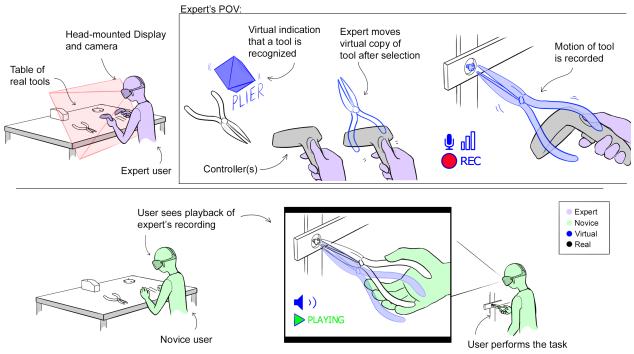

Figure 6: First, Expert Users visually scan the set of tools available within their work space once. Upon recognition, virtual models of the tools and objects are rendered from the repository and can then be controlled with the aid of the controller. This enables the experts to record instructions along with voice instructions, which can then both be replayed for novice users later.

4.3 Virtual Repository

As mentioned above, upon detection of an real object, a virtual model is rendered into the scene as an means for authors to create the instructions. However, this feature relies on existence of a pre-made virtual model repository to use of ProcessAR.

We used large CAD repositories, such as GrabCAD and TraceParts, to find pre-existing virtual models for our application. Also, it is standard practice today for stock part manufacturers (McMaster Carr) to release CAD models of their catalog. Hence, our reliance on pre-made models for standard parts is not unreasonable, especially in manufacturing, the main target application domain for ProcessAR. Custom parts can be created with additional effort through 3D scanning tools such as Qlone [50], Cognex [12] and display.land [15]. To support these needs, authors are provided with the ability to add their own repository and trained object detection data sets inside ProcessAR.

4.4 Visualization

Our approach of using a virtual placeholder leverages virtual object rendering. Instead of directly superimposing the CAD model on the physical tool upon detection, we generate a virtual marker with a tool label (orange diamond; teaser fig B) that is only replaced by the actual CAD model upon contact with the user's controller. The virtual marker has a preconfigured lifetime, which self-destructs after a certain amount of time if left untouched, thus prompting the system to re-detect the object until the user selects the correct targeted virtual tool. Once the user has made a tool selection, the system replaces the placeholder with a CAD model by searching an internal CAD database for a match corresponding to the tool's label. This approach addresses situations where temporarily inaccurate detection causes confusion with mismatched overlays of the (wrong) virtual model on a physical object.

To compensate for the difficulty of recognizing small objects, we physically grouped such objects of identical class into plastic containers, which are easier to recognize and train, and substituted the training data with the container images while keeping the same label. For example, a small screw is an object that tends to have a low training accuracy, so we grouped multiple screws in a bin and trained the bin instead. This approach reduced the difficulty of model training without posing additional distractions to the user, because small component objects placed in containers are (very) common in the context of the task. We are aware that some procedural tasks demand multiple instances of expendable objects, such as a bins of screws or two shelf brackets and multiple wood screws for installing a shelf on a wall. However, rendering all instances at once could cause confusion and generate visual clutter in the virtual space. We overcome this problem by allowing these specific virtual objects to self-duplicate at their initially rendered position after being grabbed by the user's controller. Meanwhile, reusable tools such as screwdrivers and drills were configured to only render once.

4.5 Authoring

4.5.1 3D Animation. Virtual Object Recording: Once a required tool or object is correctly identified, located, and initialized, the authoring user can interact with the virtual tool and demonstrate the "intended action" they wish to perform with the tool. This demonstration can be recorded according to the needs of the spatial task by grabbing virtual objects using the Oculus controller. The recording is achieved with a script to store the three translational coordinates for the position, four quaternion coordinates for the orientation, and corresponding time stamps. The amount of data and the difference in time between each data point varies based on frame rate of the system. The recording process is initiated with a button trigger only when the user's virtual hands are manipulating the virtual tool. This constraint was added to prevent unintended recording of multiple objects. The motion of the tool is still recorded as long as the user remains within the recording mode, which in turn enables the user to adjust and reposition the virtual tool if needed. Each recording is a self-contained unit called a checkpoint. Each checkpoint contains one 3D animation corresponding to a virtual tool. To clarify the authoring workflow, the author of the AR instruction has to interact with the virtual tool and demonstrate the action to be recorded. Then they have to pick up the real tool and perform the real action before moving to the next step.

Voice Recording: Aside from the instructional motion capture, ProcessAR also enables the user to perform voice recordings for the purpose of clarifying tasks or to explain possible error cases. This feature allows users to use voice-over during instruction capture, both for object manipulation and video creation. After each recording, an audio file is saved (in WAV format) for later use when deploying the instructions to a novice. Finally, time stamps are used to ensure motion and vocal recordings are in sync upon deployment to the novice.

4.5.2 2D Video Embedding. Video Recording: ProcessAR also supports video recording as an alternative to 3D instructions. This allows users to mix-and-match 3D animation and 2D video demonstrations, whichever they believe to be more appropriate for the current task step. The user signals the beginning and end of the video recording through a button press. During recording, instead of using controllers, the user only has to directly demonstrate the task with their physical hands. Meanwhile in the background, ProcessAR captures the real-time stereo video stream viewed by the user through the Oculus headset and stores the video clip using H.264 compression for later overlay into 3D space. Users are also allowed to import external video files if they choose not to use real-time recording.

Expert Mode - For expert users to author video instructions, ProcessAR features a novel interaction modality to overlay videos in AR. We created a radial UI panel that automatically loads the first color frame of each 2D video onto a designated panel button, associating each entry on the radial menu to a specific video. As mentioned previously, these videos can be sourced externally or recorded in-situ. The radial panel resembles a paint palette. As opposed to picking colors, the user can then freely navigate to each panel entry through the controller joystick, which prompts a virtual window to appear and enables the user to preview the clip. Once a video is chosen, the user can pull a video marker object out from the window to tag virtual replicas that are relevant to the video instruction. Similar to 3D animation recording, a set of replicas tagged by a video is considered a checkpoint. (Refer fig. 1 E & F)

Novice Mode - In novice mode, novices can view videos previously overlaid by the expert users to complete the current procedural task in progress. When the novice proceeds to a video instruction, ProcessAR invokes a checklist of objects pending detection by YOLO. The list of objects are authored by the expert user as a means of asynchronous task guidance. This functionality ensures that the novice is aware of the required tools for a task prior to taking action. Once the checklist is fulfilled, a virtual screen featuring the corresponding task video is overlaid at the final detected object's 3D position. The video screen can be freely dragged around by the users in case it is blocking their view. To play or pause the video, the user just needs to simply gaze at or look away from the screen. To alleviate the trouble of manually reorienting the screen to face the user, we applied a look-at camera matrix such that the screen is always facing the front of the Oculus HMD. This feature was added to support the perspective awareness design goal (D3). The design rationale for this feature is based on the preliminary study, where, with video-based instructions, most of the novice

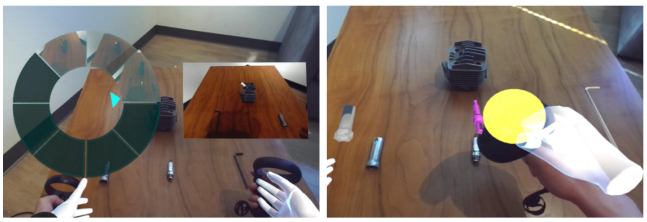

Figure 7: (Left) UI Panel for video preview and selection. (Right) Expert user tagging objects with the virtual marker to associate with the video.

Figure 8: (Left) Expert mode UI to preview, edit, delete created checkpoints. (Right) UI in novice mode, allowing the user to keep track of their progress.

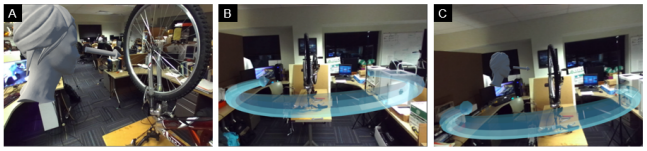

Figure 9: (A) A Virtual head indicating direction and position of where the next instruction will begin; (B) a navigational arrow rendered on the AR work space, based on the authors' movement; (C) the adaptive POV trigger and navigational arrows can be used together.

users had a tendency to pause the video mid-way during task completion to avoid cognitive overload. Our functionality offers the same kind of feature in AR.

4.5.3 User Interface. Expert/Novice User Interface: For monitoring instruction creation, ProcessAR includes a virtual control panel to help users keep track of and edit their procedure recordings. The panel follows the user's field of view in AR and can be called upon and hidden through the trigger buttons of the Oculus Touch controllers. Virtual buttons on the panel provide preview and deletion mechanisms for both user motion and voice recordings, as well as display information about accumulated recordings. The activation of object detection and the display of the spatial mapping can also be controlled via virtual toggles. Through the panel, users are empowered to freely preview and manipulate recorded instructions analogous to common video editing software.

4.5.4 ***Adaptive Point Of View Trigger & Navigation.*** Due to the 3D nature of the instructions, well-authored 3D AR animated instructions are able to holistically capture tool operation, without needing recording from multiple perspectives. This contrasts with 2D video, where often the most appropriate perspective to view the instruction has to be pre-identified by the SME and then used to capture the content. As observed in our preliminary design study, novices sometimes missed the location where the motion of the virtual 3D content was starting from. To avoid that, we were inspired by results from the literature [64] and thus provide a virtual rendering of a head with an arrow pointing towards the most appropriate perspective in our system (see fig.10 C). This "best" perspective is based on observing the position and orientation information of the HMD at the time of authoring, which also obviates the need for an explicit action from the EU to set such a trigger. Once a viewer collides their head by more than 65% volumetric overlap (threshold identified by trial and error) with such a trigger, ProcessAR activates the corresponding instruction sequence (similar to fig.10 A). The position of the POV is obtained from the HMD pose of the user. Yet, first trials identified that a major obstacle for this method was when users of different heights use the system, which makes it challenging to match a head pose.

Thus, we decided to automatically adapt the rendering of the virtual head trigger to the current users' height, by adjusting the y-axis of the head model, based on an initial calibration of the users' eye level from the HMD when the system starts (Refer fig.10 B). A focal point is set about 50 cm (average working distance between hand and eye) away along the directional normal of the gaze of the HMD, and using simple trigonometry we re-position the virtual trigger to suit the new users' height. With this method the system can automatically adapt for both tall and short users.

Navigation: Navigation enables the NU to navigate around the work space. To guide the NU to the location of the next trigger point, the system automatically renders the path for navigation by storing the position of the HMD worn by the author at the time of creation. A corresponding semi-circular arrow is then rendered by taking three points, the initial, middle, and final tracked positions as input. The semi-transparent arrow then indicates the location of the next task to the NU.

5 FINAL USER EVALUATION

We invited 12 users (7 male and 5 female) to test and evaluate features of the ProcessAR system. Four users (all male) were chosen as Expert Users (EUs) due to their prior expertise in performing car maintenance (2 users) and bike repair operations (hobby cyclists, 2 users). The other 8 users were Novice Users (NUs). The study duration was 90 minutes, and all users were compensated with a $20 Amazon gift card. All Expert studies were conducted before the studies with the Novices. For later analysis, all studies were recorded from a first-person view via VR screen mirror capture and from a third-person point of view via an external camera.

Upon arrival all users were asked to fill out a standard pre-survey questionnaire asking for background information, such as gender, age, height, level of expertise with the task, and familiarity with AR or programming. The users had an age range between 19 and 28 years. None of the users had prior experience with AR, two users had played VR games before on HTC Vive headset. The EUs were between 5' 7" to 6' 1" tall, while the novices had a height range of 5' 2" to 6'0".

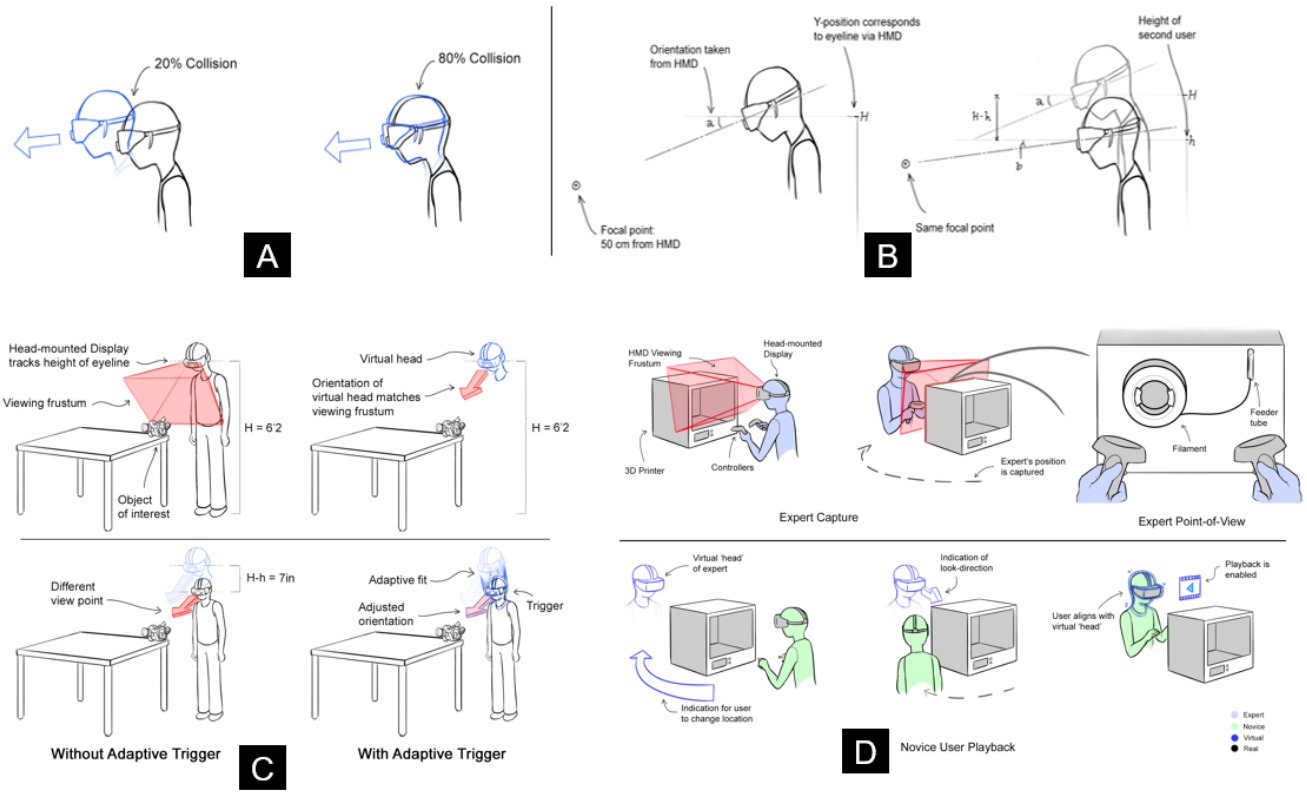

Figure 10: (A) Sketch of a volumetric collision, where a user collides with a reference virtual head, created by the expert, used by ProcessAR system as a trigger. (B) Illustration of the method ProcessAR uses to adapt the virtual head to users with different heights. (C) A representation of the adaptive POV trigger mechanism, for users with varying heights. (D) (Top) Shows a workflow of an author navigating a work space, while the novice follows the path rendered by ProcessAR based on the authors movement (D)(Bottom).

The EUs were asked to create two sets of instructions for the same task by using two variations of the AR authoring system. The features involved in both these systems are described in table 2. The "baseline" system can create both 3D and 2D instructions but lacks features such as object detection to load the virtual tools. Instead it relies on button presses to traverse through a library of virtual tools and loads them. Similarly, 2D videos can be recorded but attaching them to specific objects is not possible; instead, they are arbitrarily placed in the work space. Finally, the user's position is not tracked, and point of view triggers and navigational arrows for task guidance are not rendered. These specific features were disabled for the baseline system, as this represent reasonably closely current commercial AR authoring tools, such as by PTC [49] and Microsoft's system [39]. The other condition used by the EUs was our ProcessAR system with all features enabled.

One of the tasks to be authored was the same as in the design study (i.e., to create AR instructions for an engine assembly operation). The other task was similar to the bike repair operation described in the design study, which involved removing and reinstalling the front wheel of the bicycle. Yet, here we added the task of replacing the brake pad of the rear wheel, along with tightening the brake wire of the single pivot side-pull brake. The study was counterbalanced such that the order of the AR conditions alternated between users. That is, if one user experienced the baseline system first, the second user then used the ProcessAR system first. Following the EUs, the NUs were asked to complete the task using instructions generated by both systems (Baseline and ProcessAR), as created by the EUs, completing each task twice.

5.1 Measures

For the EUs we measured the time of completion for each instruction, while for the NUs we measured the task completion time. Both the EUs and NUs were asked to rate each system feature on a 5-point Likert scale questionnaire, followed by the SUS for evaluating the usability of the whole system, and a NASA TLX survey for perceived workload. For NUs we also investigated the number of rewind button hits, i.e., we measured how many times the Novices had to replay a particular instruction. Finally, we also collected post study feedback about the system from all users.

6 RESULTS & DISCUSSION

Compared to the baseline, the ProcessAR system shows promising results with differences in qualitative data, such as a cognitive load

Table 2: Two modes of AR authoring system and their features

Modes	3D animate	2D record	2D video embedding (to objects)	Object Recognition	POV Trigger	Navigation
-						
ProcessAR	yes	yes	yes	yes	yes	yes
Baseline	yes	yes	-	-	-	-

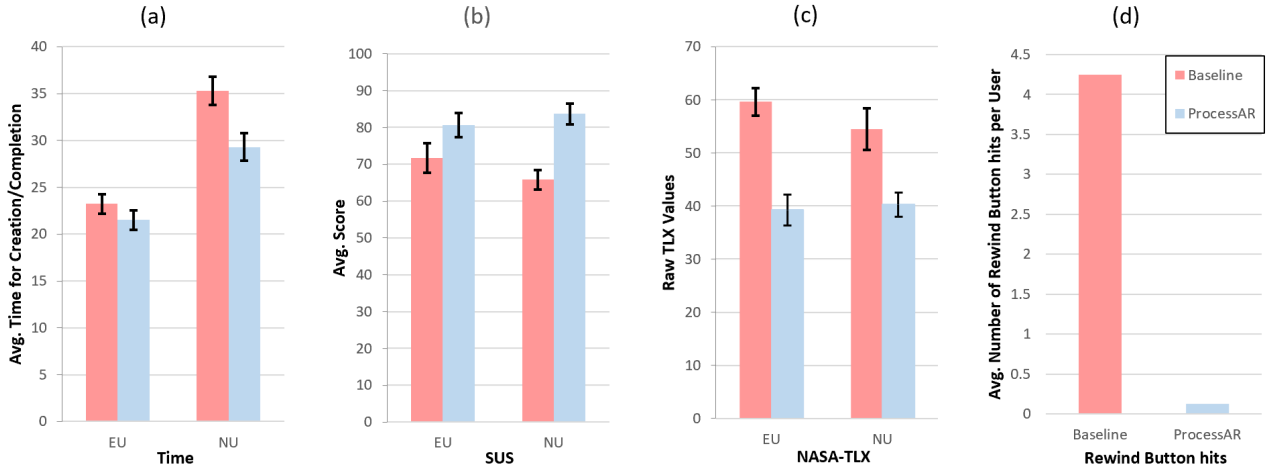

Figure 11: (a) Time taken by EU to complete instructional creation, or NU task completion time for both modes; (b) SUS scores for the systems followed by NASA TXL scores (C); (d) average rewind button hits for each user in both modes. Error bars show standard error.

reduction, as identified through NASA TLX scores and 5-point Likert scale questions capturing users' perception of the system, for both Expert and Novice users. The quantitative measures provided a reliable reference for the usability of ProcessAR. Other metrics together with user feedback provide a basis for deeper discussions.

Task time: in terms of time for instruction creation by EUs, the ProcessAR condition (M=21.5, SD=2.93) is relatively close to the Baseline condition (M=23.2, SD=3.01). The sample t-test results with Fisher's exact test are: $t(6) = 0.80, p = 0.44 > 0.05$, with effect size $r = 0.3$. Thus, there was no statistically significant difference in overall task time between both conditions. Yet, post-study video analysis and observations shows that the Experts were slightly faster at selecting the desired virtual tool by simply looking at the tool, instead of traversing through the library via button presses. This information combined with the feedback for Q2 of the EU Likert-scale questionnaire yields a better perspective. Q2 queries EUs about their agreement to the following statement "I was able to easily identify and interact with my virtual tools". Here, 75% of the EUs strongly identified with the statement for ProcessAR, while for the baseline case 50% of EUs responded with slight disagreement. While the speed of NUs was also observed to be faster (figure 11 a), we attribute this to the fact that most NUs had to rewind instructions as they had missed the beginning of the instruction. This was unnecessary in the ProcessAR condition due to the adaptive point of view trigger. This observation is further substantiated through the data from the Likert questionnaire (Q6), data on rewind button hits, and post-study feedback from users. There was only one user who had to hit the rewind button to watch an instruction again under the ProcessAR condition, while all users had to rewind at least thrice in the baseline condition. Another interesting observation was that except for one instance NUs did not perceive a need to use the controller, as all instructions were rendered when and where required, due to the adaptive point of view trigger. This led to users effectively interacting directly with the tools, without the need to constantly switch between controllers and the physical tool. Q5 of the NU Likert scale questionnaire assessed NUs agreement to "It was easy to switch between performing the task and navigating to the next instruction". Overwhelmingly all NUs rated the ProcessAR condition as being easy, while 6 out of 8 NUs strongly disagreed with the statement for the baseline. This observation is best corroborated by the following comment of one of the NUs during the post-survey, where they stated: "*I would have been fine with using the controller to move to the next step, had I not known about the hands-free [controller-less] option. That was cool.*"

Usability and Cognitive Load: The scores from the SUS questionnaire indicate no statistically significant difference in terms of usability between ProcessAR and baseline for expert users (EUs). The scores for ProcessAR (M=80.6, SD=6.6) and the baseline

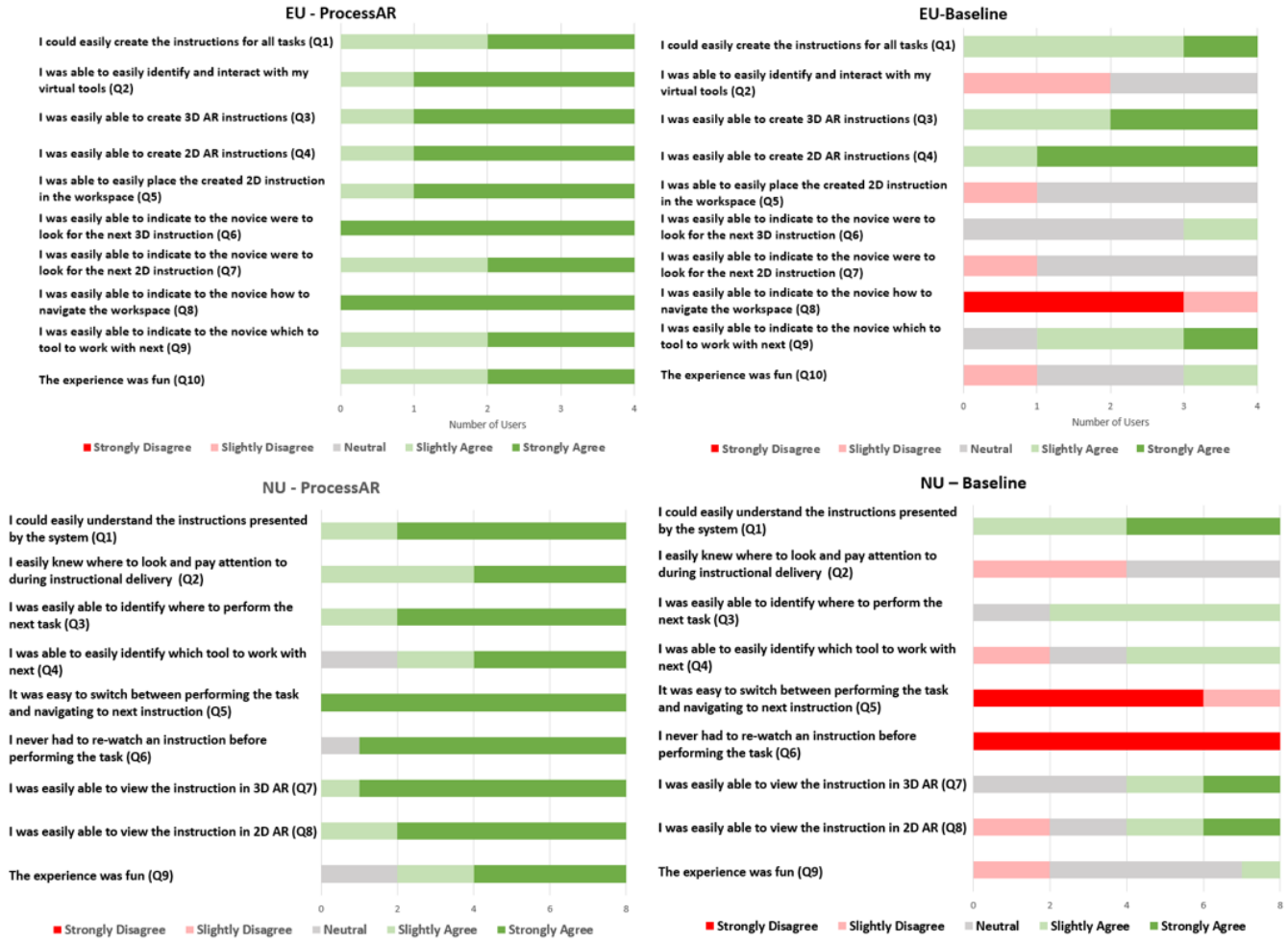

Figure 12: The Likert-scale statements used for evaluation along with an illustration of the responses for each statement.

condition (M=71.7, SD=8.2) do not differ significantly according to a t-test: $t(6) = 1.60$, $p = 0.14 > 0.05$, $r = 0.54$. Yet, the average SUS score of 80.6 for ProcessAR is encouraging, as an average score of 70 and above translates to "excellent" usability, as indicated by analysis of Bangor et al. [2]. However, for NUs there was a statistical difference in SUS scores between ProcessAR (M=83.7, SD=7.9) and the baseline condition (M=65.8,SD=7.4): $t(14) = 4.67$, $p = 0.0004 < 0.05$, $r = 0.78$. Pairing these results with the Likert-scale questionnaire and post-study feedback, we find evidence that the NUs found the system to be effective for training. Thus we can state that the usability of ProcessAR was enhanced through the added features, such as the adaptive POV triggers and hands-free interaction.

The perceived workload measure in the NASA TLX survey provided better results for the ProcessAR system by both EUs and NUs. The EUs raw NASA TLX scores were (M=39.25,SD=5.83) for ProcessAR and (M=59.62,SD=5.09) for the baseline condition. The sample t-test results identified a significant difference: $t(6) = 5.19$,$p = 0.0020 < 0.05$, $r = 0.9$. For NUs the raw NASA TLX scores were (M=40.25,SD=6.36) for ProcessAR and (M=54.5,SD=11.09) for the baseline condition. The sample t-test again identified a significant difference: $t(14) = 3.15$,$p = 0.0071 < 0.05$, $r = 0.64$. Both for Experts and Novice users the results demonstrate a statistically significant reduction in workload with ProcessAR.

Finally, we would like to mention two suggestions by users. Some experts users felt the reliance on controllers to be intrusive to a seamless interaction during authoring 3D instructions. That is, the EUs had to first demonstrate the interaction with the tool/object with the virtual model via a controller, then switch to using the physical tool/object to perform the action. This is best described by a user quote "*I wish I can just show what I have to do [with the physical tool] and the 3D instructions are generated, instead of switching back and forth with the controller and the tool.*". The other feedback was "*The graphics of me holding the wrench [tool/object] could be better, like right now my virtual hands go inside the wrench for me to grab it. It would be better if the graphics was more precise*". This user was referring to the fact that to interact with the virtual tool, the controller has to collide with the virtual model and the fact that dynamic tool interaction such as the movement of pliers when

squeezing the handles or a wrench adjustment animation currently cannot be captured. Both these suggestions are acknowledged as limitations and will be pursued as part of future work.

7 LIMITATIONS AND FUTURE WORK

To improve reliability in tracking and issues due to occlusion, we avoided relying purely on computer vision and instead preferred the use of the Oculus touch controllers, which can be robustly tracked in real time (also due to their inertial measurement units, IMUs). Their reliability makes our system perform smoothly and enables the authors to create the 3D animation. Unfortunately, using the controllers occupies the users' hands and can be intrusive while performing the task. The expert is thus required to keep switching between the controllers and using their hands. If a different interaction device, such as a tracked glove [22, 70], would be employed, it might be possible to track the tool position without compromising the use of hands. Still, ProcessAR provides hands-free interaction for Novices, so this issue affected only the Expert users.

Some dynamic tools, such as measuring tapes or socket wrenches with changeable heads are hard to animate. Enabling the corresponding virtual tools to be dynamic and being able to record interactions with them automatically would enables even better tool operation tutorials to be produced in AR. Finally, a feature which could detect task completion or incorrect performance, similar to work such as Drill Sergeant [58], would also be beneficial.

Future work on ProcessAR could expand on the navigational feature, to create and enable use of more complex navigational maps. It is also theoretically possible to create AR content with ProcessAR and to then deploy the instructional content on any other AR-enabled device, such as a phone or a tablet. Such an implementation would expand the reach of AR to a wider audience while still ensuring ease of instructional content creation for the expert authors.

8 CONCLUSION

We presented ProcessAR, an AR authoring platform with a robust implementation for procedurally developing 3D and 2D AR instructions, especially for delivering spatio-temporal instructions. Currently, most industries still use the inefficient method of one-on-one training to train their employees for tasks such as machine operation and manufacturing. Unfortunately, due to the large number of retiring workers [13] there is an increased demand for newer and more flexible workers to meet the challenges of the future where many technical changes leave a skills gap [29]. Our system also extends current AR-based instruction systems, such as Microsoft Dynamics 365 Guide, by supporting custom animations of 3D models of tools. Our user evaluations with experts and novices corroborated that SMEs are easily able to use our system to author AR instructions with no experience in AR programming and only minimal training on the tool. Based on the outcomes of our work, we can conclusively state that ProcessAR allows experts to create task tutorials more quickly than other forms of instruction. Moreover, ProcessAR can be used to train workers on demand, which has the potential to speed up manufacturing workforce skill transfer, make it less expensive and enables more flexibility to re-allocate workers across workflows.

Until recent advances in commercially viable head mounted devices for AR systems, there was no technology sufficiently reliable to capture and deliver spatio-temporal aspects of human knowledge similar to what we demonstrated here. With the advent of AR this has been rectified but this technology comes with the price of instructional content being hard(er) to create. We believe systems such as ProcessAR enable and encourage more SMEs to create instructions and share their hands-on skills and knowledge, enabling them to encapsulate their experience for the benefit of a wider audience and thus also make training scalable across a variety of tasks and procedures.

ACKNOWLEDGMENTS

We would like to thank Vedant Kadam, Wentao Zhong, Rutvik Mehta, Sai Swarup Reddy and Matthew David Rumple for their help and support during the development and evaluation of the system. We also thank Eugene Lee for his help with the sketches used in this publication.

This work was partially supported by U.S. National Science Foundation awards FW-HTF 1839971 (http://www.nsf.org/), OIA 1937036 and OIA 2033615. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

REFERENCES

[1] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. Association for Computing Machinery, New York, NY, USA, 311–320. https://doi.org/10.1145/2501988.2502045

[2] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[3] James Baumeister, Seung Youb Ssin, Neven AM ElSayed, Jillian Dorrian, David P Webb, James A Walsh, Timothy M Simon, Andrew Irlitti, Ross T Smith, Mark Kohler, et al. 2017. Cognitive cost of using augmented reality displays. *IEEE transactions on visualization and computer graphics* 23, 11 (2017), 2378–2388.

[4] Bhaskar Bhattacharya and Eliot H Winer. 2019. Augmented reality via expert demonstration authoring (AREDA). *Computers in Industry* 105 (2019), 61–79.

[5] Boeing. 2019. Boeing: The Boeing Company. Retrived June 2, 2019 from https://www.boeing.com/.

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] Techsmith Camtasia. 2019. Camtasia: Screen Recorder and Video Editor. Retrieved September 18, 2019, from https://www.techsmith.com/video-editor.html.

[8] Yuanzhi Cao, Xun Qian, Tianyi Wang, Rachel Lee, Ke Huo, and Karthik Ramani. 2020. An Exploratory Study of Augmented Reality Presence for Tutoring Machine Tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376688

[9] Scott Carter, Pernilla Qvarfordt, Matthew Cooper, and Ville Mäkelä. 2015. Creating tutorials with web-based authoring and heads-up capture. *IEEE Pervasive Computing* 14, 3 (2015), 44–52.

[10] Pei-Yu Chi, Joyce Liu, Jason Linder, Mira Dontcheva, Wilmot Li, and Bjoern Hartmann. 2013. Democut: generating concise instructional videos for physical demonstrations. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. Association for Computing Machinery, New York, NY, USA, 141–150. https://doi.org/10.1145/2501988.2502052

[11] Pei-Yu Peggy Chi, Daniel Vogel, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2016. Authoring illustrations of human movements by iterative physical demonstration. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 809–820. https://doi.org/10.1145/2984511.2984559

[12] cognex. 2020. cognex. Retrieved Feb 5,2021, from https://www.cognex.com/products/machine-vision/3d-machine-vision-systems/in-sight-3d-l4000.

[13] D'Vera Cohn and Paul Taylor. 2010. Baby Boomers Approach 65 - Glumly. Retrived September 15, 2019 from https://www.pewresearch.org/fact-tank/2010/12/29/baby-boomers-retire/.

[14] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. 2014. You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video.. In *BMVC*, Vol. 2. BMVA Press, Nottingham, UK, 3.

[15] display.land. 2020. display.land. Retrieved May 5,2020, from https://get.display.land/.

[16] Daniel Eckhoff, Christian Sandor, Christian Lins, Ulrich Eck, Denis Kalkofen, and Andreas Hein. 2018. TutAR: augmented reality tutorials for hands-only procedures. In *Proceedings of the 16th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*. Association for Computing Machinery, New York, NY, USA, 8. https://doi.org/10.1145/3284398.3284399

[17] Carmine Elvezio, Mengu Sukan, Ohan Oda, Steven Feiner, and Barbara Tversky. 2017. Remote collaboration in AR and VR using virtual replicas. In *ACM SIGGRAPH 2017 VR Village*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3089269.3089281

[18] Michele Fiorentino, Rafael Radkowski, Christian Stritzke, Antonio E Uva, and Giuseppe Monno. 2013. Design review of CAD assemblies using bimanual natural interface. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 7, 4 (2013), 249–260.

[19] Markus Funk, Mareike Kritzler, and Florian Michahelles. 2017. HoloCollab: a shared virtual platform for physical assembly training using spatially-aware head-mounted displays. In *Proceedings of the Seventh International Conference on the Internet of Things*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3131542.3131559

[20] Markus Funk, Lars Lischke, Sven Mayer, Alireza Sahami Shirazi, and Albrecht Schmidt. 2018. Teach Me How! Interactive Assembly Instructions Using Demonstration and In-Situ Projection. In *Assistive Augmentation*. Springer, Cham, 49–73.

[21] GIMP. 2019. GIMP: GNU Image Manipulation Program. Retrieved September 18, 2019, from https://www.gimp.org.

[22] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. 2019. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 41. https://doi.org/10.1145/3306346.3322957

[23] Michihiko Goto, Yuko Uematsu, Hideo Saito, Shuji Senda, and Akihiko Iketani. 2010. Task support system by displaying instructional video onto AR workspace. In *2010 IEEE International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, Los Alamitos, CA, USA, 83–90.

[24] Ankit Gupta, Dieter Fox, Brian Curless, and Michael Cohen. 2012. DuploTrack: a real-time system for authoring and guiding duplo block assembly. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. Association for Computing Machinery, New York, NY, USA, 389–402. https://doi.org/10.1145/2380116.2380167

[25] Taejin Ha and Woontack Woo. 2007. Graphical tangible user interface for a ar authoring tool in product design environment. In *International Symposium on Ubiquitous VR*. Citeseer, Gwangju, Korea, 1.

[26] Matthias Haringer and Holger T Regenbrecht. 2002. A pragmatic approach to augmented reality authoring. In *Proceedings. International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, Los Alamitos, CA, USA, 237–245.

[27] Steven Henderson and Steven Feiner. 2010. Exploring the benefits of augmented reality documentation for maintenance and repair. *IEEE transactions on visualization and computer graphics* 17, 10 (2010), 1355–1368.

[28] Inkscape. 2019. Inkscape: Draw Freely. Retrieved September 18, 2019, from https://inkscape.org.

[29] Deloitte Insights. 2018. Deloitte skills gap and future of work in manufacturing study. Retrieved September 18, 2019, from https://www2.deloitte.com/content/dam/insights/us/articles/4736_2018-Deloitte-skills-gap-FoW-manufacturing/DI_2018-Deloitte-skills-gap-FoW-manufacturing-study.pdf.

[30] SAE International. 2019. Formula SAE. Retrieved September 18, 2019, fromhttps://www.sae.org.

[31] Jarrod Knibbe, Tovi Grossman, and George Fitzmaurice. 2015. Smart makerspace: An immersive instructional space for physical tasks. In *Proceedings of the 2015 International Conference on Interactive Tabletops & Surfaces*. Association for Computing Machinery, New York, NY, USA, 83–92. https://doi.org/10.1145/2817721.2817741

[32] Taihei Kojima, Atsushi Hiyama, Takahiro Miura, and Michitaka Hirose. 2014. Training archived physical skill through immersive virtual environment. In *International Conference on Human Interface and the Management of Information*. Springer, Cham, 51–58.

[33] Thomas Kubitza and Albrecht Schmidt. 2015. Towards a toolkit for the rapid creation of smart environments. In *International Symposium on End User Development*. Springer, Cham, 230–235.

[34] Suha Kwak, Woonhyun Nam, Bohyung Han, and Joon Hee Han. 2011. Learning occlusion with likelihoods for visual tracking. In *2011 International Conference on Computer Vision*. IEEE Computer Society, Los Alamitos, CA, USA, 1551–1558.

[35] Michael Laielli, James Smith, Giscard Biamby, Trevor Darrell, and Bjoern Hartmann. 2019. LabelAR: A Spatial Guidance Interface for Fast Computer Vision Image Collection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 987–998. https://doi.org/10.1145/3332165.3347927

[36] David F Lohman. 1996. Spatial ability and g. *Human abilities: Their nature and measurement* 97 (1996), 116.

[37] Michael R Marner, Andrew Irlitti, and Bruce H Thomas. 2013. Improving procedural task performance with augmented reality annotations. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, Los Alamitos, CA, USA, 39–48.

[38] Lockheed Martin. 2019. ow Lockheed Martin is Using Augmented Reality in Aerospace Manufacturing. Retrived from https://www.engineering.com/AdvancedManufacturing/ArticleID/19450.

[39] Microsoft. 2020. Overview of Dynamics 365 Guides. Retrieved May 5,2020, from https://docs.microsoft.com/en-us/dynamics365/mixed-reality/guides/.

[40] Peter Mohr, Bernhard Kerbl, Michael Donoser, Dieter Schmalstieg, and Denis Kalkofen. 2015. Retargeting technical documentation to augmented reality. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 3337–3346. https://doi.org/10.1145/2702123.2702490

[41] Peter Mohr, David Mandl, Markus Tatzgern, Eduardo Veas, Dieter Schmalstieg, and Denis Kalkofen. 2017. Retargeting video tutorials showing tools with surface contact to augmented reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 6547–6558. https://doi.org/10.1145/3025453.3025688

[42] Lars Müller, Ilhan Aslan, and Lucas Krüßen. 2013. GuideMe: A mobile augmented reality system to display user manuals for home appliances. In *International Conference on Advances in Computer Entertainment Technology*. Springer, Cham, 152–167.

[43] Michael Nebeling, Janet Nebeling, Ao Yu, and Rob Rumble. 2018. ProtoAR: Rapid Physical-Digital Prototyping of Mobile Augmented Reality Applications. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173927

[44] Oculus. 2019. Oculus Rift Store. Retrieved September 18, 2019, from https://www.oculus.com/experiences/rift.

[45] Oculus. 2019. Oculus Software Development Kit. Retrieved September 18, 2019, from https://developer.oculus.com.

[46] Ohan Oda, Carmine Elvezio, Mengu Sukan, Steven Feiner, and Barbara Tversky. 2015. Virtual replicas for remote assistance in virtual and augmented reality. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. Association for Computing Machinery, New York, NY, USA, 405–415. https://doi.org/10.1145/2807442.2807497

[47] SK Ong and ZB Wang. 2011. Augmented assembly technologies based on 3D bare-hand interaction. *CIRP annals* 60, 1 (2011), 1–4.

[48] Nils Petersen and Didier Stricker. 2012. Learning task structure from video examples for workflow tracking and authoring. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, Los Alamitos, CA, USA, 237–246.

[49] PTC. 2019. Vuforia Expert Capture. Retrieved May 5,2020, from https://www.ptc.com/en/products/augmented-reality/vuforia-expert-capture.

[50] qlone. 2020. qlone. Retrieved May 5,2020, from https://www.qlone.pro/.

[51] Rafael Radkowski and Christian Stritzke. 2012. Interactive hand gesture-based assembly for augmented reality applications. In *Proceedings of the 2012 International Conference on Advances in Computer-Human Interactions*. Citeseer, Valencia, Spain, 303–308.

[52] Vijaimukund Raghavan, Jose Molineros, and Rajeev Sharma. 1999. Interactive evaluation of assembly sequences using augmented reality. *IEEE Transactions on Robotics and Automation* 15, 3 (1999), 435–449.

[53] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE Computer Society, Los Alamitos, CA, USA, 779–788.

[54] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE Computer Society, Los Alamitos, CA, USA, 7263–7271.

[55] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. (2018).

[56] Dirk Reiners, Didier Stricker, Gudrun Klinker, and Stefan Müller. 1999. Augmented reality for construction tasks: Doorlock assembly. *Proc. Ieee And Acm Iwar* 98, 1 (1999), 31–46.

[57] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive Body-Driven Graphics for Augmented Video Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 622. https://doi.org/10.1145/3290605.3300852

[58] Eldon Schoop, Michelle Nguyen, Daniel Lim, Valkyrie Savage, Sean Follmer, and Björn Hartmann. 2016. Drill Sergeant: Supporting physical construction projects through an ecosystem of augmented tools. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*.

Association for Computing Machinery, New York, NY, USA, 1607–1614. https://doi.org/10.1145/2851581.2892429

[59] Rajeev Sharma and Jose Molineros. 1997. Computer vision-based augmented reality for guiding manual assembly. *Presence: Teleoperators & Virtual Environments* 6, 3 (1997), 292–317.

[60] Autodesk Sketchbook. 2019. Autodesk Sketchbook. Retrieved September 18, 2019, from https://sketchbook.com.

[61] Enox Software. 2019. OpenCV for Unity. Retrieved September 18, 2019, from https://enoxsoftware.com/opencvforunity.

[62] Maximilian Speicher, Kristina Tenhaft, Simon Heinen, and Harry Handorf. 2015. Enabling industry 4.0 with holobuilder. In *INFORMATIK 2015*, Douglas W. Cunningham, Petra Hofstedt, Klaus Meer, and Ingo Schmitt (Eds.). Gesellschaft für Informatik e.V., Bonn, 1561–1575.

[63] Stereolabs. 2019. Stereolabs ZED - Unity Plugin. Retrieved September 18, 2019, from https://github.com/stereolabs/zed-unity.

[64] Mengu Sukan, Carmine Elvezio, Ohan Oda, Steven Feiner, and Barbara Tversky. 2014. Parafrustum: Visualization techniques for guiding a user to a constrained set of viewing positions and orientations. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. Association for Computing Machinery, New York, NY, USA, 331–340. https://doi.org/10.1145/2642918.2647417

[65] Anna Syberfeldt, Oscar Danielsson, Magnus Holm, and Lihui Wang. 2015. Visual assembling guidance using augmented reality. *Procedia Manufacturing* 1 (2015), 98–109.

[66] Richard Tang, Xing-Dong Yang, Scott Bateman, Joaquim Jorge, and Anthony Tang. 2015. Physio@ Home: Exploring visual guidance and feedback techniques for physiotherapy exercises. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 4123–4132. https://doi.org/10.1145/2702123.2702401

[67] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 161–174. https://doi.org/10.1145/3332165.3347872

[68] Jonathan Wai, David Lubinski, and Camilla P Benbow. 2009. Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology* 101, 4 (2009), 817.

[69] Matt Whitlock, George Fitzmaurice, Tovi Grossman, and Justin Matejka. 2020. AuthAR: Concurrent Authoring of Tutorials for AR Assembly Guidance. In *Graphics Interface*. CHCCS/SCDHM, University of Toronto,Ontario, Canada, 431 – 439.

[70] Sang Ho Yoon, Ansh Verma, Kylie Peppler, and Karthik Ramani. 2015. HandiMate: exploring a modular robotics kit for animating crafted toys. In *Proceedings of the 14th International Conference on Interaction Design and Children*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/2771839.2771841