Pham, Truong An; Moesgen, Tim; Siltanen, Sanni; Bergstrom, Joanna; Xiao, Yu

ARiana: Augmented Reality based In-Situ Annotation of Assembly Videos

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2022.3216015

Published: 01/01/2022

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 September 2022, accepted 10 October 2022, date of publication 19 October 2022, date of current version 27 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3216015

RESEARCH ARTICLE

ARiana: Augmented Reality Based In-Situ Annotation of Assembly Videos

TRUONG AN PHAM[1], TIM MOESGEN[1], SANNI SILTANEN[2,3], JOANNA BERGSTRÖM[4], AND YU XIAO[1]

[1]Department of Communications and Networking, School of Electrical Engineering, Aalto University, 02150 Espoo, Finland
[2]Dimecc Ltd., 33100 Tampere, Finland
[3]Department of Computing Sciences, Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland
[4]Department of Computer Science, University of Copenhagen, 1017 Copenhagen, Denmark

Corresponding author: Truong An Pham (truong.pham@aalto.fi)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Annotated videos are commonly produced for documenting assembly and maintenance processes in the manufacturing industry. However, according to a semi-structured interview we conducted with industrial experts, the current process of creating annotated assembly videos, in which the annotator annotates the video capturing the expert's demonstration of assembly and maintenance process, is cumbersome and time-consuming. The key challenges include three key problems in annotation: (1) unnecessary extra communications between field workers and annotators, (2) lack of suitable camera gear, and (3) wasting time in the manual removal of non-informative portions of captured videos. Because annotation always follows video capture, the problem 1 remains out of scope for state-of-the-art video annotation tools. And making the assumption of a perfect captured video, which covers no occlusion and contains only relevant assembly or maintenance information, causes problem 2 and 3. As a result, we have developed ARiana, a wearable augmented reality-based in-situ video annotation tool that guides field experts to create annotations efficiently while conducting the assembly or maintenance tasks. ARiana has three key features that include context-awareness enabled by hand-object interaction, multimodal interaction for annotation on the fly, and real-time audiovisual guidance enabled by edge offloading. We have implemented ARiana on Android-based smart glasses, equipped with first-person camera and microphone. In a usability test based on attempting to assemble a toy model and to annotate the recorded video simultaneously, ARiana demonstrated higher efficiency and effectiveness compared to one of the state-of-the-art video annotation tools, in which the assembling process is followed by the annotation process. In particular, ARiana helps users finish annotation tasks four times faster, and increase the annotation accuracy by 23%.

INDEX TERMS Augmented reality, first-person videos, multimodal interaction, process documentation, video annotation, workflow extraction.

I. INTRODUCTION

Knowledge preservation and transfer are increasingly pressing problems in work across domains. According to [1], expertise and experience are two of main losses of a company when an employee retirement process is executed. An economy blog titled *What if an expert retired?* [2] puts this in numbers: it mentions a report of a loss of over 27000 years of experience due to the retirement of 700 experts. Besides retirement, the high turnover rate and the trend towards digitization and automation of industrial work have further increased the need for efficient knowledge preservation and transfer.

Video is considered one of the richest mediums for capturing activities or processes. It has been increasingly used for documenting and demonstrating industrial operations.

The associate editor coordinating the review of this manuscript and approving it for publication was Giuseppe Desolda.

Compared with traditional face-to-face training, video-based training saves time of training supervisors and allows learners to move at their own pace [3]. However, certain information (e.g. the amount of force to apply, the operating temperature, or the product model number) may be missing or difficult to extract from videos even with the state-of-the-art computer vision techniques. Therefore, videos captured by cameras need to be annotated with instructions and hints before they can be used for training employees [4], [5].

Since automatic image or video annotation [6], [7], [8] is still in its infancy, manual labeling through crowdsourcing [9], [10] has been the most common way of producing annotated visual data on a large scale. The drawback of crowdsourcing-based annotation [11], [12], [13] arises from the fact that the annotators are typically amateurs and have little knowledge regarding the video contents, which leaves the domain-specific, personalized, and expertise-based knowledge inaccessible.

This paper describes ARiana, a wearable augmented reality-based (AR-based) video annotation tool that allows field workers with domain knowledge to efficiently create high-quality annotated videos on site. To define design goals of the tool, we first explore the key problems in the current practice of creating annotated video in the manufacturing industry through a literature survey and a formative interview with ten industrial experts from an international elevator company.

According to our study, current problems are mainly related to setups of camera gear (e.g., occlusions caused by third-person camera setups or frustration due to the wearing of bulky cameras during maintenance work) and manual efforts required for creating annotations (e.g., filtering out irrelevant content from video or communicating with field workers to create or refine descriptions of operations in the videos in case annotations are made afterwards by others). The communications costs for discussing the content of the video are removed automatically when the videos are recorded and annotated by the same person, who is the operator of the assembly and maintenance process in this case. This confirms benefits of integrating the video annotation process into the assembly or maintenance process, which also includes the process of video recording, instead of separating them as is practiced in the conventional approach.

To address these problems, we propose three key features of ARiana, including context awareness enabled by hand-object interaction, multimodal interaction for annotation on the fly, and real-time audiovisual guidance. With these features, ARiana has proven to accomplish the following three design goals according to a usability test with 18 participants.

1) Minimising the time and effort wasted in identifying and removing non-informative parts of the recorded video.
2) Providing high-quality annotation with minimal cognitive overhead.
3) Providing real-time (i.e. low latency) feedback in the video annotation process.

The results of the usability test showed that ARiana is easy to use for annotating three types of workflow information, including the starting and ending points, associated objects (e.g. tools, materials, machinery components) and instruction for each work step. Compared with Ajalon [14], a state-of-the-art video annotation tool that supports semi-automatic work step segmentation and object association list extraction, ARiana guided participants to annotate workflow information at least 60% faster without inflating the level of experienced cognitive load in the annotation process. Moreover, the quality of object annotation is significantly higher with ARiana than with Ajalon.

To the best of our knowledge, ARiana is the first AR-based in-situ annotation tool for industrial assembly and maintenance work. Although it has been built based on existing techniques such as object detection, speech recognition, and gesture recognition, it has gone beyond the state-of-the-art and has created the following scientific contributions.

1) Context awareness based on hand-object interaction to automatically identify the associated objects used to complete a work step, to remove redundant video segments, and to eliminate noninformative video segments by idle stage detection.
2) A novel multimodal interface that uses speech and pointing gestures to label objects, improving the accuracy of annotating associated objects.
3) An edge-offloading-based architecture design to support real-time guidance for on-the-fly video annotation process.

The rest of this paper is structured as follows. We introduce the background and the related work in Section 2. Section 3 presents the current issues in the existing video annotation tools. The design goals for a new video annotation tool are defined in Section 4. Section 5 describes the design and development of two main components of ARiana, including context awareness and multi-modal interaction. The system evaluation including a comparison with Ajalon [14] is presented in Section 6. We discuss the implications of the design goals and the findings from the evaluation in Section 7 before we conclude the work in Section 8.

II. BACKGROUND AND RELATED WORK

In this section, we first introduce different granularity and types of annotations for assembly and maintenance videos. Different features or functions are often needed to implement different types of annotations.[1] Next, we describe the landscape of different types of video annotation tools. The analysis of the advantages and disadvantages of the existing tools has provided the background for our studies of the current challenges in the video annotation process as well as for derivation of the design goals for new tools, including ARiana.

[1]Note that a maintenance process in practice consists of assembly and/or disassembly operations. In the rest of this paper, we choose maintenance as an example to study the video annotation solution that would work for both assembly and maintenance process.

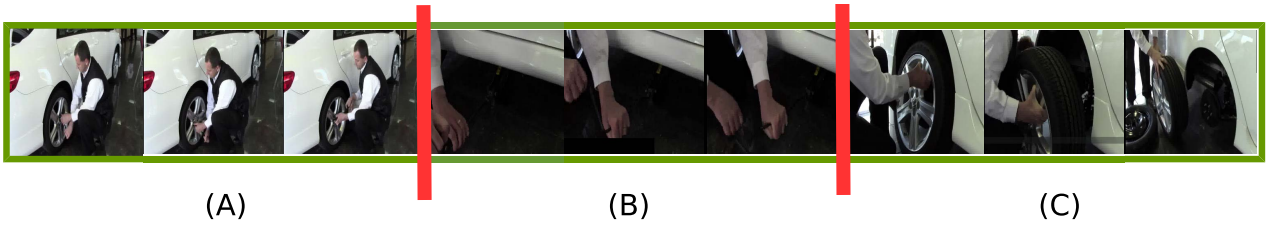


FIGURE 1. Three successive work steps in a how-do video of a tire-changing process [15]: (A) Loosening the screws in the middle of the wheel, (B) jacking up the car, (C) taking out the wheel. The starting and ending points of the second step are indicated with red vertical lines.

A. GRANULARITY AND TYPES OF VIDEO ANNOTATIONS

Video can be annotated at different granularity level, depending on the intended use of video annotations. For example, a how-to video or a segment of it can be annotated with a sentence that describes the task demonstrated in the video or the segment [17], [18]. Compared with such video-wise or segment-wise annotations, annotations with a finer granularity, such as bounding boxes outlining objects of interest [11], [19], [20] or even pixel-wise object annotations [21], [22], are needed for training machine learning models for computer vision tasks such as object recognition. We summarize the following types of annotations that are desired for assembly videos below, based on a literature survey on instructional videos (e.g., [18], [23]).

Starting and ending points of each work step. Assume that a video captures the whole assembly or maintenance process demonstrated by a field expert step by step. Given such a video, annotators are expected to divide the video into segments, with each segment representing a work step. As shown in Figure 1, given a piece of video that contains three steps of a tire changing process, annotations can be added to the corresponding frames that represent the starting and ending point of each segment (i.e., work step). Instead of work step segmentation, separation points are defined because the ending of one step may not necessarily be the starting of the next step due to irrelevant content in between. For example, workers take a rest or leave to find tools.

Object association list. After segmenting a video into work steps, the next step is to understand the operations in each step. To describe the operations, it is necessary to know which tools, materials, or components are needed for completing the operations. This is typically implemented using object recognition techniques. Each video segment representing a work step can be annotated with an object association list. An example is provided in Figure 2. In the list, the locations of relevant objects (e.g., tools, machinery parts, or materials) in each video frame are annotated in the format of bounding box. Note that three-dimensional (3D) locations of objects in the working environment may be calculated and annotated using 3D object localization techniques, such as simultaneous localization and mapping (SLAM) [24], [25], and marker-based tracking [26], [27]. However, due to the low robustness of these techniques in industrial environment, we have left the annotation of 3D locations for future work, and have focused on locations of objects in two-dimensional (2D) images in this work. Besides locations, the objects may be annotated with names in text, 2D images, or even 3D models of the objects [27], [28].

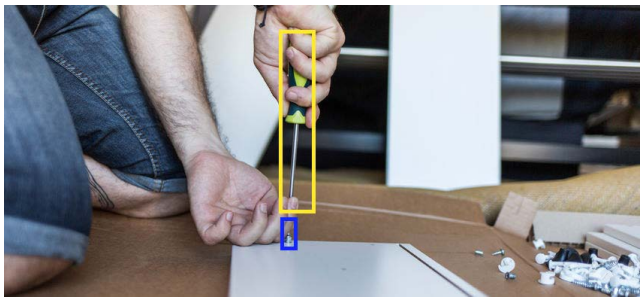


FIGURE 2. The tool, a screwdriver, is annotated with a yellow bounding box, while the assembly component, which is a screw in this case, is annotated with a blue bounding box. The bounding boxes indicate the locations of the objects in question in the image [16].

Instruction. The objects association list alone is not sufficient to explain the operations in each work step. In how-to video sets [15], [17], each work step is commonly described with a sentence in human-written or human-spoken language. For example, in [15], ''Start by loosening each bolt'' is annotated for the the bolt-loosening work step in the car tire replacement video. Such description can be considered as instruction for completing the work step. It contains both objects (e.g. names of objects) and verbs (e.g. actions). In some cases, the instruction also includes tips or tricks to avoid common mistakes, for example a recommendation to apply moderate force when screwing into the wood to avoid breaking. We have focused on text-based instructions in this work, although the instructions may also be provided in the form of audio, video, or animation [26].

B. VIDEO ANNOTATION TOOLS

Many video annotation tools have been developed and used since 1990s, for instance in creating instructional videos [4], [29], analyzing video experiments [30], sharing comments on videos through social networks [31], and generating labeled data for training machine learning models [14]. These tools are typically used for annotation after the video has been captured. In contrast to these tools, our proposed solution, ARiana, focuses on annotating videos on the fly.

From the interaction perspective, the mouse and keyboard were the most commonly used input devices before tablets and smartphones with touchscreen become popular. In case of

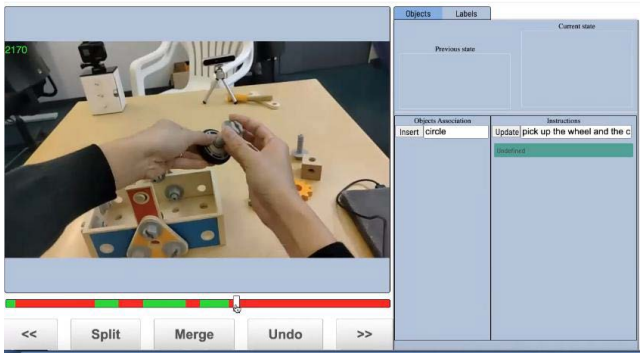


FIGURE 3. Ajalon's interface consists of three major components, including split and merge buttons for segmenting work steps, an object association box for inputting names of an associated object immediately after labeling it on screen view, and instruction box for providing text-based instructions.

annotating pre-recorded video, it is preferred to use a mouse and a keyboard to achieve precise pointing and fast typing, although users have reported that they become fatigued when using these devices to annotate lengthy video [14]. When users need to use their hands to conduct other tasks simultaneously, hands-free interaction with the annotation tool becomes an essential requirement. Early work has been conducted on speech interaction [32] or gaze-based [33]. For example, speech interaction and mouse interaction were combined in [32] to achieve a significantly faster annotating speed. On the other hand, in [34], the authors observed that the inconsistent eye-tracking of the sensor made gaze-based interaction unreliable in mixed reality based annotation application. In this work, we propose to combine speech interaction with gestural interaction to enable effortless in-situ annotation of first-person videos during maintenance field work.

Instead of designing a novel interaction model to improve user experience, several video annotation tools [14], [35], [36] have focused on developing intelligent features that can reduce manual effort and speed up the annotation process. For example, the video annotation tool presented in [35] used automatic object segmentation to save 78% of manual effort, although the results of the object segmentation algorithms still needed to be manually corrected occasionally. Similarly, the authors in [36] make the annotating process 11.3% faster by using automatic gesture segmentation and recognition for annotating video events. For assebmly video annotation, the recently published tool Ajalon [14] advanced further to support automatic detection of objects of interest as well as work step segmentation. As illustrated in Figure 3, Ajalon provides a web-based interface for users to edit the results of work step segmentation and object detection.

Similarly to Ajalon, ARiana applies computer vision techniques to enable real-time work step and object detection. In addition, ARiana is a wearable cognitive assistance application that provides context-aware guidance for video annotation during field work, with the aim of improving annotation performance with minimal cognitive overhead. In other words, ARiana focuses on multi-tasking scenarios (i.e., annotation as a secondary task in parallel to the primary task of assembly or maintenance operation) and addresses challenges related to the use of wearable interfaces like smart glasses.

Previous interaction design for smart glasses [37], [38], [39], [40] has not focused on multitasking scenarios and therefore cannot be applied directly in our work, while other works on interaction design for multitasking scenarios have not addressed wearable devices. For example, Liu et al. [41] proposed a novel way of taking notes while watching learning videos online. Their design focused on the achieved learning outcomes, and the length and appropriateness of the taken notes; however, it did not consider the cognitive load of users. Compared with previous wearable cognitive assistance systems like Jarvis and Gabriel, ARiana involves the same concept of edge offloading that Gabriel does but extends the support from single-tasking to multi-tasking.

Regarding the quality and productivity of video annotation, previous works [9], [30], [42], [43] have focused on impacting factors from the perspective of the annotators, ignoring the impact from the quality of the recorded video [18], [23], [44], [45]. In the design of ARiana, we have considered the video recording process and its impact on the quality and efficiency of video annotation.

III. CURRENT ISSUES

To identify the key problems and challenges in the current practice of video annotation in the manufacturing industry, we conducted a literature survey and a formative interview at an international elevator company. In this section, we present the implementation of the interview and the problems that were identified, with references from the literature.

A. PARTICIPANTS

At the elevator company, field experts often demonstrate maintenance methods and record the demonstration with video cameras for training purposes. The videos are annotated afterwards by a technical documentation team. To gain a better understanding about the annotation process and the tools in use, we invited five field experts with a role of maintenance-method developer and five annotators from the technical documentation team to join the interview. The results of our interview, presented below, also confirmed the need for involving both parties in the study.

B. INTERVIEW PROCEDURE

The formative interview was conducted as a group interview with all 10 participants. The participants were interviewed in a neutral environment, a meeting room in this case, instead of at the maintenance work site, to promote innovative thinking about their work processes as well as to accommodate the annotators. The interview lasted one hour.

The participants were first informed about the purpose of the interview which was to investigate both sides of the annotation process and to discover issues that are experienced

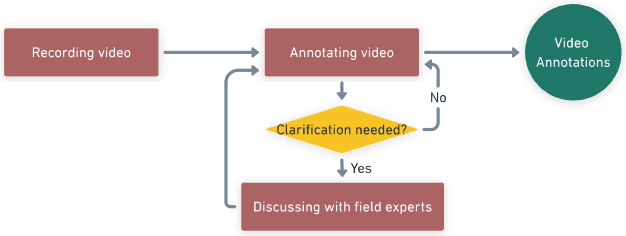
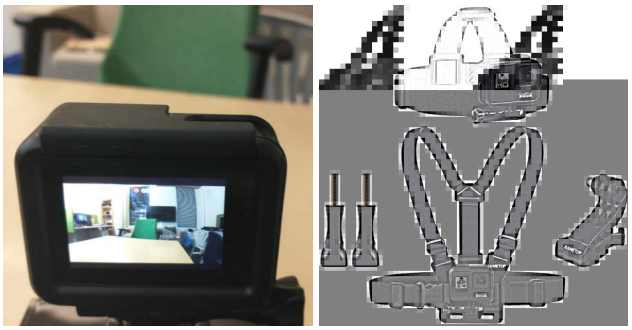


FIGURE 4. The current process of creating annotated videos of maintenance processes in an elevator company. Video is captured with chest cameras during maintenance processes. Video annotation involves adding description of the maintenance processes to the corresponding parts of the video. If the video annotator is not able to produce exact annotations for any part of the video, due to occlusions or lack of domain knowledge, discussion with field experts who produced the videos is arranged.



(a) (b)

FIGURE 5. Camera setups used in the elevator company for recordings demonstrative videos. (a) A typical third-person camera setup. (b) First-person camera setup.

on either side during the process. They were instructed to freely answer, discuss, and complement one another's observations about the annotation process. The interviewer presented questions about the phases of the annotation process, following a semi-structured interview protocol. The questions are included in Appendix A. The interviewer engaged with the participants in the discussion and took notes of key observations regarding the annotation process as well as direct references to the problems experienced in that process.

C. PROBLEMS IN THE CURRENT ANNOTATION PROCESS

After the interview, the interviewer summarized the discussions based on the observation and collected notes. Combined with the results of a literature survey, the following three problems were identified:

1) PROBLEM 1: EXTRA COMMUNICATIONS BETWEEN FIELD WORKERS AND ANNOTATORS

As described in Figure 4, both the video annotator and the field expert must sit together for re-checking, discussing, and annotating ambiguous parts in the video. Therefore, the video annotation process is often delayed due to repetitive discussions. As a result, time of both the field expert and the annotator is consumed. Two out of the five field experts in our interviews expressed feelings of annoyance in discussing events that occurred in the video afterwards. This problem occurs not only in annotation of elevator maintenance works but also in other domain-specific video annotation, such as that of surgical videos in [46].

2) PROBLEM 2: LACK OF SUITABLE CAMERA GEARS

In the context of maintenance fieldwork, we found three disadvantages of third-person camera setups (as illustrated in Figure 5a, including occlusion, space requirement, and safety risk. These are detailed in the following accounts: 1) All video annotators considered occluded camera viewpoint a frequent and serious problem in the annotation process. Information loss due to occlusions may prevent the annotators from understanding events that occurred in a video. However, due to the limited working space, it is hard to position a third-person camera distant enough to capture the entire work area with no occlusion. 2) As one field expert pointed out, the working environment in an elevator shaft requires a high safety level. For example, it is absolutely forbidden to climb on a balustrade to mount a camera, and the walls in the shaft are made mainly of concrete or other materials on which camera mounting is impossible. In summary, field experts and video annotators would prefer first-person cameras that are safe to use, easy to set up, and can be placed in such a way that the issue of occlusion would be minimized. Moreover, first-person camera setups can be more suitable for capturing hand-object interactions and for detecting the worker's attention [47], [48], as people tend to focus on the objects being operated.

There are two types of camera setups for capturing first-person videos. One setup is to attach cameras to different parts of the body, such as forehead and chest as shown in Figure 5b. Adjusting the positions of cameras to make sure they capture the operations with minimal occlusions often takes considerable time. As reported in [50], users may also feel uncomfortable to wear such cameras. The other type of setup is to use the cameras embedded in augmented or mixed reality headsets, such as Microsoft Hololens, Google Glasses, or Vuzix smart glasses. However, these glasses either are so bulky that they impact the health and movement of the user [53] or do not provide a sufficiently wide field of view. Therefore, it has been challenging to find proper camera gear that is comfortable to wear and would not disturb the field work.

3) PROBLEM 3: MANUAL EFFORTS FOR REMOVING NON-INFORMATIVE PORTIONS OF CAPTURED VIDEOS

As a common procedure, field experts or annotators must analyze the video afterward to filter out non-informative portions. For example, one field expert stated, "The camera records video continuously without pausing while the field expert is doing irrelevant stuff, such as finding tools in another room, taking a phone call, or drinking coffee during the break time. As a result, the length of recorded video is significantly longer than expected." To highlight that point,

another field expert shared the experience of recording a one-hour video but retaining with only 15 minutes of content that was informative and could be annotated. This aligns with findings from a study of the open video dataset COIN [17], which consists of 11,827 videos related to 180 different tasks. For most videos in the COIN dataset, the proportion of informative parts within each video is not higher than 50%.

IV. DESIGN GOALS

To address the three problems identified in Section III, it became necessary to invent new approaches and tools for annotating the videos of maintenance fieldwork, which would also require changes to the current process of video collection and annotation in the manufacturing industry. In this work, we have explored the feasibility of in-situ annotation, which integrates video annotation into the process of field work. In other words, the annotations are collected while recording the videos, and the field experts who conduct the maintenance work play the role of annotators as well. This removes the communications costs mentioned in Section III.C by default. However, field experts may not know how to annotate a video or demonstrate a task. Therefore, we aimed to design a tool to solve the other problems mentioned in Section III and to fill the gap.

As discussed in Section III.C, first-person cameras are more suitable for capturing hand-object interactions than third-person cameras, given that they are comfortable to wear. With the rapid development of wearable technology, we envision that lightweight augmented or mixed reality glasses with embedded wide-angle lenses and microphones will become more widely available soon.

Therefore, we designed ARiana with the assumption that it will run on such glasses and use a head-mounted display to provide visual feedback for users. This design choice was made due to the following considerations. First, the cameras embedded in smart glasses allow ARiana to capture first-person video in a convenient and safe way. The video can be viewed in real time on the display of the glasses, which helps users to adjust the camera setup when necessary to ensure that operations and working environments are captured as desired. Second, annotations can be visualized as virtual content on top of related physical objects in the camera view. This allows annotators to review and edit the annotations while conducting the assembly or maintenance tasks. For the time being, we have used Vuzix M400 smart glasses for experimentation. Therefore, gaze-based interaction is not an option in designing the interface of ARiana.

To guide field workers in creating high-quality annotations in an efficient manner, ARiana has been designed to fulfill the following design goals (DGs).

DG 1 - Minimise the time and effort spent in identifying and removing non-informative parts of recorded video. To solve problem 3 mentioned in Section III.C, one solution is to collect videos in a way that does not record breaks between work steps or any other irrelevant activities. This would require either hints given directly by users or algorithms that consider the relevance of activities in real time.

DG 2 - Provide high-quality annotation with minimal cognitive overhead. The increase in cognitive load of field workers is a potential risk of multitasking required by in-situ video annotation because the worker needs to keep in mind when and how to annotate the video. Higher cognitive load may slow down the primary task and may cause errors. Therefore, it is essential that the tool can remind users of required annotations at the proper time in a natural way. This requires context-awareness and multimodal interaction to reduce cognitive load and to improve the productivity of annotation. Here, the productivity can be measured by the number of annotations, the time spent, and the number of annotation errors.

DG 3 - Provide real-time audiovisual guidance. Low latency is a common requirement for interactive AR applications, as high latency would significantly degrade the user experience. Here, latency refers to the time taken from capturing the video, processing the video, and giving feedback to the user. It consists of both processing delay and data transmission delay, if applicable.

V. SYSTEM DESIGN

In this section, we describe the detailed design of ARiana and explain how it fulfills the design goals defined in Section IV.

A. CONTEXT AWARENESS

Video annotation tools typically work in a passive way in that they simply provide an interface for users to create annotations (e.g., drawing bounding boxes and adding textual descriptions) but do not give hints or instructions regarding what and when to annotate. In this case, the quality and quantity of annotations depend on how active the users are and how accurate the user input is. We define this condition as a user-initiated annotation process. On the other hand, video annotation tools with awareness of context can be more active in collecting annotations by reminding and guiding users to create annotations. We call this concept context-aware guided annotation. ARiana is designed to support both working modes, as illustrated in Figure 6.

ARiana enables context awareness by implementing three features, including a hand-object interaction inference, associated-object detection, and speech recognition. With context-awareness enabled, ARiana can propose annotations online (e.g., bounding boxes of objects of interest or the starting and ending of each work step) and ask users to confirm or correct them. In this way, it can automatically remove the non-informative portions of videos, including breaks between work steps, to save time and effort (DG1). Moreover, it can remind users when and what to annotate (DG2). In general, instruction for each work step can be recorded while executing the work step. The relevant objects can be annotated at the beginning of the whole maintenance process (Option 1), at the beginning of each work step (Option 2), or when they appear (Option 3). We surveyed instructional video data

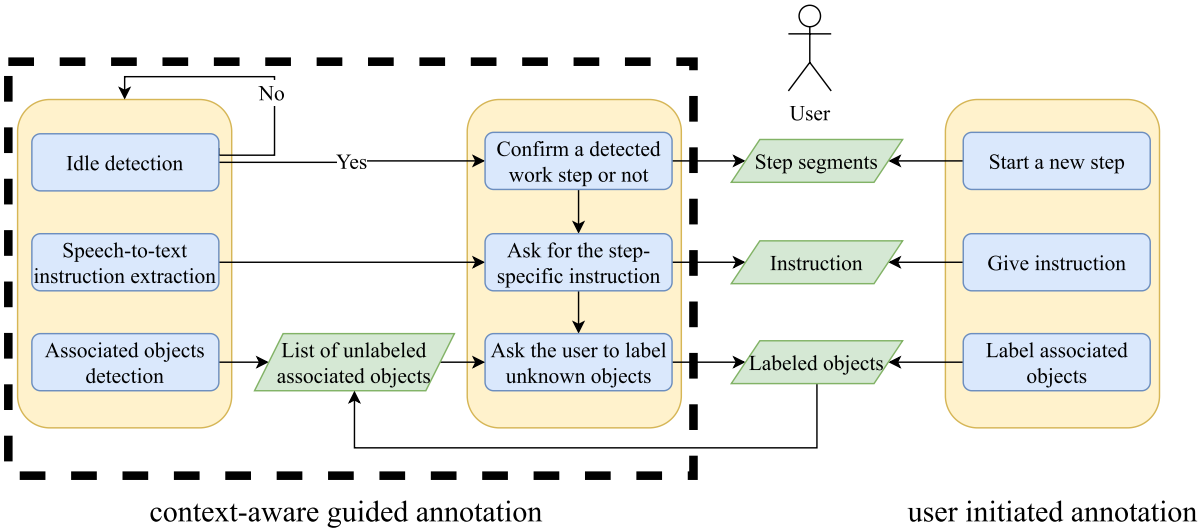


FIGURE 6. System overview of ARiana. ARiana supports two working modes, including user initiated annotation and context-aware guided annotation.

sets [17], [18] and found that it is common for experts to introduce the relevant objects at the beginning of each step (Option 2). This approach is more feasible for multi-step processes, concerning memorability. Therefore, we choose to follow Option 2 in the design of ARiana.

1) HAND-OBJECT INTERACTION

ARiana continuously monitors hand-object interaction. If a work step has not been completed if there is no break in terms of hand-object interaction. The objects refer to those declared by the user as relevant tools or materials at the beginning of each work step. The process enters an idle mode when no hand-object interaction has been detected for a period of time (typically 2 seconds, which was empirically chosen based on preliminary studies for an action in hand-gesture recognition [49]). Hand-object interaction detected after entering an idle mode indicates the starting of a new work step.

For learning to recognise hand-object interaction and detect regions of interactions, deep neural networks (DNN) such as FasterRCNN [50] or a hand-object holding detection method [51] are dominant approaches. However, DNN methods often suffer from the issue of overfitting. To avoid overfitting, we have used semantic information, including hand regions and object proposals, to detect hand-object interaction in the following three steps. Firstly, we have used the mediapipe framework [52] for implementing the detection of bare hand regions. Hand detection in the mediapipe framework consists of a palm detection model and a hand landmark model. The palm detection model trained with an in-the-wild data set [52] can achieve up to 95.7% accuracy. The hand landmark model has been trained on a real-world data set [52]. Secondly, object regions is detected by the SaliencyMDC algorithm [53] mentioned below. Thirdly, we calculate the overlaps between the regions of hands and those of objects. If the overlap exceeds 30% of a proposal box, a hand-interaction is detected. When an overlap between a hand and an object is detected in a video frame, we assume that the hand is interacting with the object in question.

2) ASSOCIATED OBJECTS

ARiana automatically detects the objects being operated by hands based on hand detection and object-proposal detection. This is the outcome of the hand-object interaction above where the interacted objects are recognized as associated objects.All the objects operated in a work step form a list of associated objects for the work step.

In ARiana, object-proposal detection is implemented with the SaliencyMDC algorithm [53], which yields a short response delay. In addition, the objects of interest are tracked with the KCF tracking algorithm [54] from the time the hand overlaps with the object's box.

With associated-object detection, it becomes possible for ARiana to cue users to annotate objects, as illustrated in Figure 7b. Moreover, because the detection algorithm automatically produces bounding boxes for the objects of interest, the user does not need to draw bounding boxes manually and can simply give names to the objects through speech input.

B. MULTIMODAL INTERACTION

ARiana is designed to provide a multimodal interface that allows users to easily and precisely annotate videos using gestures and speech during fieldwork.

To design the multimodal interaction model for ARiana, we first conducted an elicitation study. An implementation plan was subsequently designed based on the results of the elicitation study.

1) ELICITATION STUDY

One way to explore the interaction model is to allow participants to think about their preferences, such as preferred hand

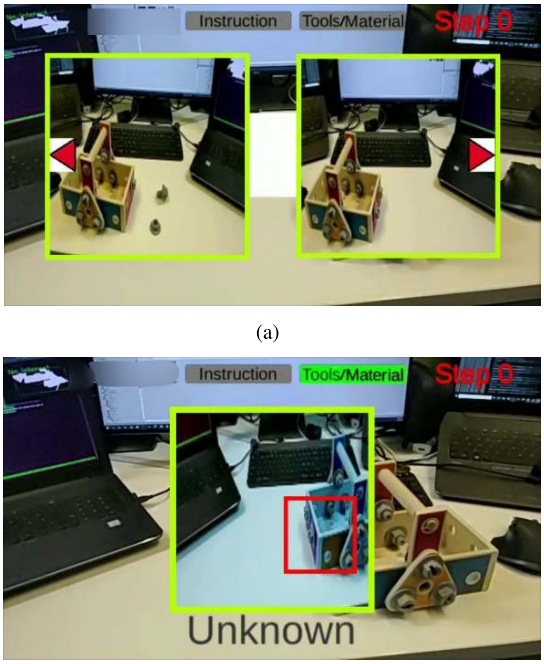


FIGURE 7. Context-aware guidance for annotation. (a) When an idle mode is detected, ARiana asks if the user has ended the step with the starting frame and the ending frame inside green boxes presented on the display. (b) When an associated object is detected, ARiana asks for the name of the object.

gestures, as done in [55]. This method is time-consuming and not effective if the participants have no prior knowledge about the tool to be developed. In this study, we conducted a literature review regarding mid-air gestures and voice commands. A list of mid-air gestures used in previous works [56], [57], [58], [59], [60], [61], [62], [63], [64] was selected based on the context of indoor working. In terms of voice interface design, there are no common standard commands; voice commands can include words, sentences, or conversations [65]. To determine suitable voice commands, we began with simple word commands instead of sentences.

a: SETUP

Eight postgraduate students with backgrounds in computer science and electrical engineering joined the user study. None of them had previous experience with smart glasses. The study was carried out in a laboratory environment simulating the experience of the assembly and maintenance processes. As illustrated in Figure 8, the test was arranged for one participant at a time. Before the test started, the participant was instructed to sit on a chair and to put on Vuzix M400 smart glasses. After that, the participant was introduced to the purpose of the study by a facilitator. The study contained three parts. The first part was to select preferred gestures and voice commands, the second one was to determine whether gesture or voice commands were more suitable for the different annotation tasks listed in Table 1, and the final part was to assemble a toy and use those gesture or voice commands for annotation.

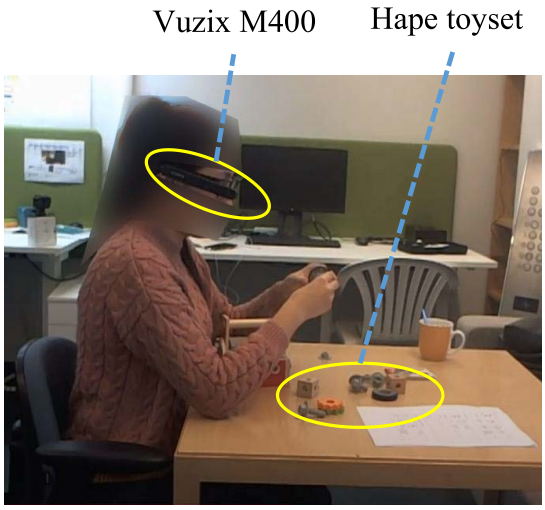


FIGURE 8. Setup of the elicitation study. A participant is performing assembly while executing the tool commands for annotation.

TABLE 1. List of annotating tasks.

Category	Annotating task
Power on/off	1. Start annotating 2. Stop annotating
Object association list annotation	3. Present an object 4. Name an object
Work step segments annotation	5. Start a new step
Instruction annotation	6. Start to give instruction
Error fixing	7. Undo the latest annotating event

b: SELECTED GESTURES AND VOICE COMMANDS

When the test began, the facilitator introduced the list of hand gestures shown in Figure 9 and voice commands as shown in Table 2 for use in creating annotations. Next, the participant would attempt all those gestures and voice commands and then select the ones they preferred or propose new ones they thought would be suitable for implementing the annotation tasks.

c: GESTURE VERSUS VOICE COMMANDS

After selecting gesture and voice commands, each participant was asked to assemble a toy set model (see Figure 10a) and to create annotations at the same time. They were instructed to complete the annotation tasks in two sessions using gestures and voice commands, respectively. They answered a survey after each session and were interviewed with open-ended questions after two sessions. The answers were used to analyze the ease-of-use and memorability of different commands. We also measured the agreement rate for each selected command following Eq. 1 proposed by Vatavu et al [66].

$$AR(r) = \frac{|P|}{|P|-1} \sum_{P_i \subseteq P} \left(\frac{|P_i|}{|P|} \right)^2 - \frac{1}{|P|-1}, \qquad (1)$$

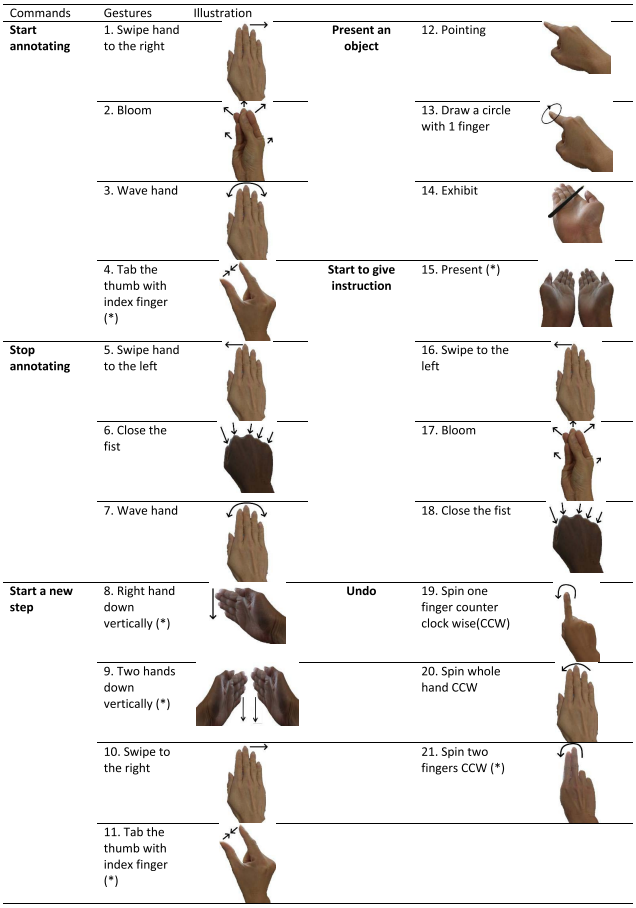


FIGURE 9. List of hand gestures for the elicitation study of ARiana's interaction design. (*) indicate gestures proposed by participants.

TABLE 2. List of voice commands.

Annotating tasks	Voice commands
1. Start annotating	Start
	Open
	Begin
2. Stop annotating	Close
	Finish
	End
	Shutdown
3. Start a new step	Next step
	New step
	Start a new step
4. Start to give instruction	Instruction
	Insert instruction
	Add instruction
	New instruction
	Next instruction
	The instruction is
	The instruction for this step is
5. Undo the latest annotating event	Undo
	Remove
	Cancel

where r represents the selected command, P represents the set of command selections, and P_i represents the set of command selections for command i.

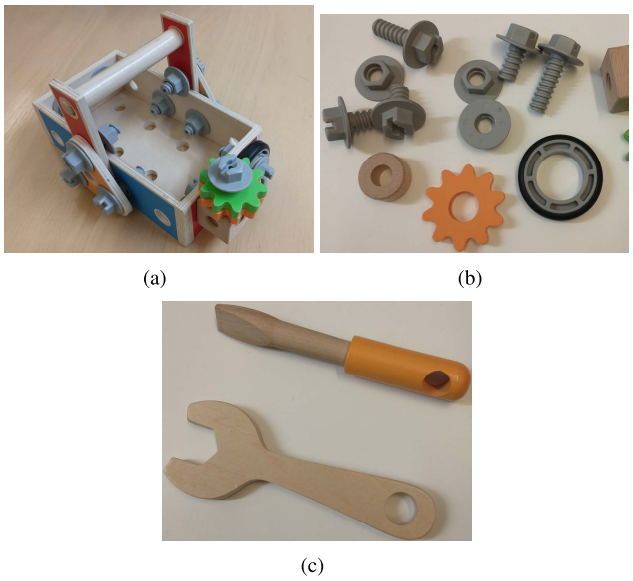


FIGURE 10. HAPE toy set with (a), (b) components and (c) tools.

Figure 11a shows that the difference between gesture and voice commands was small from the ease-of-use perspective. The advantage of voice commands was more obvious in terms of memorability, as shown in Figure 11b. Because hands were used for implementing both the primary and secondary tasks in case of in-situ video annotation, the memorability of gesture commands appeared to decrease. This was confirmed by six participants.

Regarding agreement rates, we defined three levels of agreement, from low to medium to very high. As shown in Figure 12, voice commands dominated all six annotating tasks listed. According to the study results, we chose voice commands for all the annotating tasks except selection of objects for annotation. In case of object annotation, the ease-of-use score for gesture-based object annotation was 4.75, while that of memorability was 4.875. In addition, five participants highly recommended using pointing gestures for selecting objects. Therefore, we chose to use pointing gestures for locating objects and speech for annotating object names. For this reason, the objects association list is not listed in Figure 11a or Figure 12.

2) IMPLEMENTATION OF THE INTERACTION MODEL

The proposed interaction model consists of a voice interface and a mid-air-gesture interaction interface. We have implemented four voice commands using the IBM Watson speech-to-text cloud service [67]. These commands include start annotation, stop annotation, start a new step, and start to give instruction.

We have applied a vision-based approach for recognizing the pointing gestures and localizing the pointing position in two dimensions to locate designated objects, as shown in Figure 13. Two components of the pointing gesture recognition algorithm are hand-skeleton detection as in

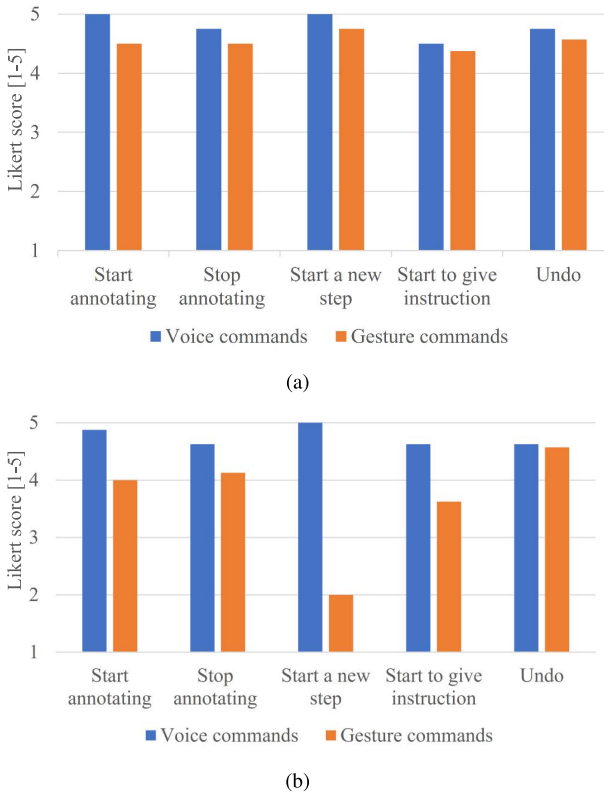


FIGURE 11. Voice versus gesture commands. (a) Ease-of-use. (b) Memorability.

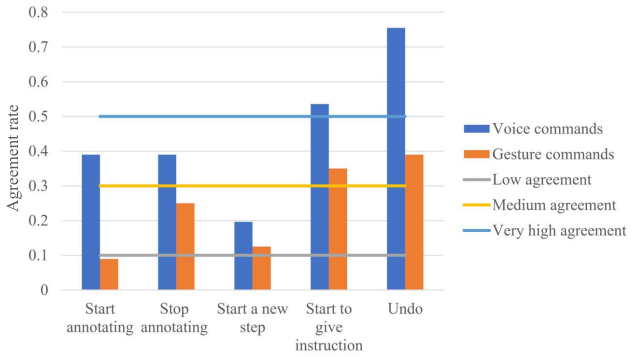


FIGURE 12. Agreement rate of gesture commands versus voice commands. The three target lines represent three agreement levels in previous interaction elicitation studies [55].

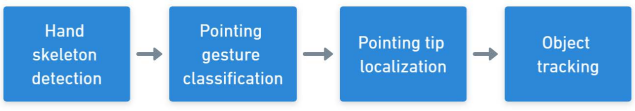


FIGURE 13. Pointing gesture localization algorithm.

Figure 15 and pointing-gesture classification. To implement the hand-skeleton detection, we have used the mediapipe framework [52] to take advantage of fast-speed orientated implementation of the hand skeleton. To increase the speed of pointing-gesture classification, we use a thresholding technique for classifying pointing gestures based on the accumulated angles between joints, which are extracted from the heat maps of the hand skeleton. The accumulated angles between hand joints can be expressed following Eq. 2.

$$AcuAngles = \sum_{j_i \subset J}^{j_1} \arccos(\angle \overrightarrow{j_i j_{i+1}} \overrightarrow{j_{i+1} j_{i+2}}), \tag{2}$$

where J is the set of 3D locations of hand joints and j_i is the 3D location of hand joint i. To classify the pointing gesture, ARiana only needs to check whether $AcuAngles$ is within a range of $[\alpha, \beta]$

According to our best knowledge in prototyping voice-based interaction, speech recognition accuracy and speed is a key factor to impacting significantly on the user experience [65]. To achieve low latency and high accuracy, we have used the IBM Watson speech recognition service, which is implemented with DeepSpeech [68], to implement the speech recognition module for ARiana. In contrast to video data, audio streaming is rapid due to lightweight data. Moreover, IBM Watson has trained their machine learning model for speech recognition with a huge dataset; thus, it is one of the most reliable state-of-the-art speech recognition platforms. However, even if a highly accurate speech recognition algorithm can be used in the future, noisy environments and recognition of words spoken by surrounding people must be considered as a drawback. Noise-filtering algorithms or context-aware speech recognition may enhance the speech recognition result. However, in this study, we assumed that we have a qualified speech recognition that work accurately in the noisy industrial environment.

C. EDGE OFFLOADING

Although high-end smart glasses such as the Vuzix M400 model already use powerful mobile chip sets (e.g., the Snapdragon XR1: 8 Kryo-385 cores, graphics processing unit [GPU]: Adreno 630), they are still not sufficiently powerful to support real-time video analytics with a high frame rate. For example, Vuzix M400 glasses can execute the hand-skeleton detection and tracking [52], [69] tracking [52], [69] at 9.26 frames per second (fps) and object proposals detection at 10 fps. Offloading heavy computation from mobile devices such as smart glasses to more powerful computers has been widely used for shortening processing delays [70], [71]. In practice, the communication overhead caused by data transmission from smart glasses to remote computers must be considered when choosing the resources for data processing. Concerning the low latency requirement of wearable AR applications (DG 3), we propose to run computation-intensive computer vision algorithms at the edge of the Internet (e.g., base stations), closer to where the smart glasses are located than the remote cloud. In cases in which there is no Internet connection, the edge can be a portable computer (e.g., an HP Z VR Backpack v2 which is equipped with Intel® Core i7, NVIDIA GeForce RTX

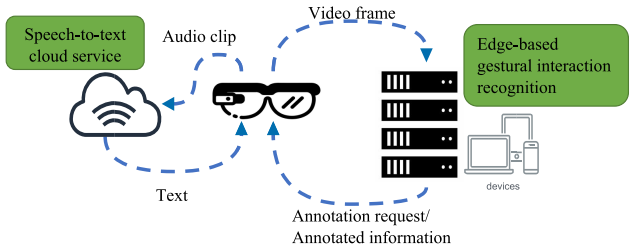


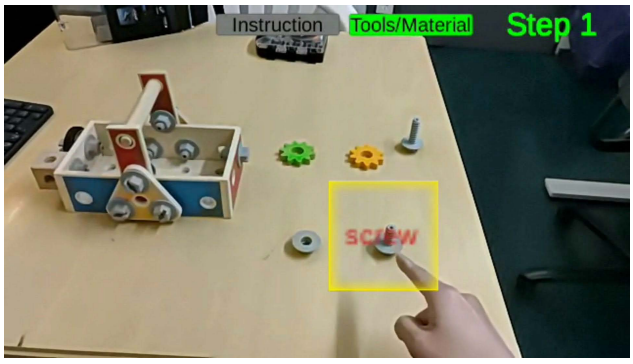
FIGURE 14. An example of a deployment scenario. Annotation requests are generated in the context-aware guided annotation mode. Annotated information is generated in the user-initiated annotation mode.

2080 and Realtec ac) connected to the smart glasses via a wireless local area network. An example of a deployment strategy for ARiana is shown in Figure 14. In this study, hand-object interaction detection and pointing-gesture recognition are deployed on powerful GPU edge server to optimise the performance. In contrast, the speech recognition processor is deployed on the cloud because the IBM Watson cloud service is used for speech recognition, and the recognition response is fast enough to capture the speech of users.

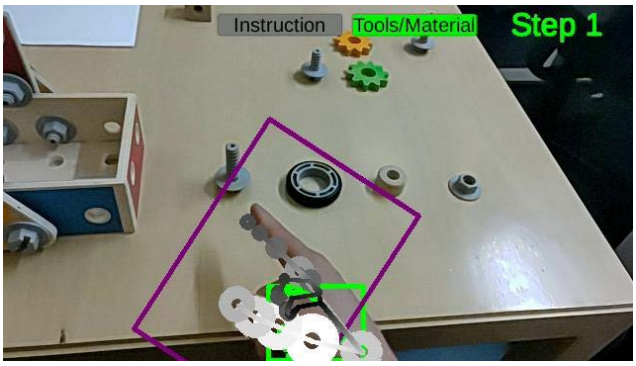
D. THE VISUAL INTERFACE OF ARiana

The visual interface of ARiana is shown in Figure 15. It highlights the work step index and shows two buttons (i.e., instruction, tools or materials) on the top. If the user is recording an instruction, or annotating tools or materials, the corresponding button is highlighted in green. If the user is providing an instruction or annotating tools or materials, the corresponding button is highlighted in green. As illustrated in Figure 7a, ARiana asks the user to confirm the automatically detected starting and ending points of a work step. With speech recognition, ARiana can automatically collect spoken instructions given by the field expert during the assembly or maintenance process. In addition, with automatic detection of work steps, ARiana can request users to record instructions in case no speech has been detected during the work step. Such a design can remind the user of the next annotation step and therefore reduce cognitive load.

As shown in Figure 15a, a designated object is overlaid with a yellow bounding box and its label. There are two problems that may occurs in this labeling interaction. First, the label may be provided inaccurately due to a mistake by the user. Second, the bounding box may be placed in an incorrect position due to the inattention of the user while pointing. Therefore, instead of being visualized in only a few frames, the bounding box and label overlay on the objects remain for a maximum of 5 seconds or until the tracking is lost. This feature should allow users to check and fix the labeling. In practice, the pointing gesture is recognised at first. Next, the pointed object is localized. Finally, the KCF object tracking [54] is used for tracking the pointed object on the fly. With tracking, the bounding box and the label can follow the objects that the user annotates.



(a)



(b)

FIGURE 15. First-person view recording using ARiana. a) The graphical user interface for the user. The interface consists of three core elements including a see-through camera view as background, an annotation information bar on the top, and the bounding box with labeled text. b) The debugging interface of ARiana. In this interface, the extracted hand skeleton is visualized with circles for finger joints and lines for finger pose. The purple box represents the bounding box around the detected hand.

VI. USABILITY TEST

We conducted a user study to test the usability of ARiana and evaluate whether the design goals have been accomplished. Moreover, we compared ARiana with Ajalon, a state-of-the-art AI-powered tool for creating annotations of pre-recorded assembly videos, in terms of quality of annotations, annotation performance, and cognitive loads of users. Like ARiana, Ajalon also supports automatic object tracking and work-step detection. This section describes the study design, implementation, and the evaluation metrics of the study and analyzes its results.

A. EVALUATION MEASURES

To evaluate how well the design of ARiana had met the three DGs, we defined corresponding evaluation measures for each one. These measures are listed below and summarized in Table 3.

1) DG1

ARiana was expected to automatically remove non-informative segments, such as the breaks between work steps, from the recorded videos. We defined idle duration as the period during which the field worker was not conducting any

TABLE 3. Summary of evaluation measurements and results.

Design goal	Sub Goals	Measures	Results
DG1	Reduce annotation time	Savings factor A	0.6 - 8.4
		Savings factor B	0.6 - 5.2
	Automatically remove non-informative parts	Percentage of idle duration in the total duration of the recorded video	2.8% (>60% half amount of videos in COIN dataset [17])
DG2	Quality of annotation	Occlusion rate of annotated objects	4.7%
		Annotation error rate	39.7% of 234 annotations
	Ease-of-use	Questionnaire (5-point Likert scale)	3.62 (Overall) 4.17 (Work-step segmentation) 2.59 (Objects association list) 4.11 (Instruction)
	Cognitive load	NASA-TLX (7-point Likert scale)	3.3 (Mental demand) 2.3 (Physical demand) 2.3 (Temporal demand) 3.2 (Performance) 3.8 (Effort) 3.4 (Frustration)
DG3	Low latency response	Latency	1.46s (Smartglass only) 0.85s (Cloud-based) 0.68s (Edge-based)
		User satisfaction regarding the delay in response (5-point Likert scale)	3.7 (Work-step segmentation) 2.9 (Object association list) 4.2 (Instruction)

maintenance or annotation task and calculated the percentage of idle duration in the total duration of the recorded video.

2) DG2

ARiana was expected to guide field workers to create high-quality annotations with minimal cognitive overhead. Regarding the quality of annotation, we measured the amount of collected annotations and computed the ratio of incorrect annotations to the total number of manual annotations. Ambiguity negatively affects the knowledge transfer process because it prevents the human or machine learning algorithms from understanding the knowledge correctly [72]. In case of object annotations, we considered an annotation as correct if a bounding box contained only one object and the object was not occluded by any other object. Examples of correct and incorrect object annotations are presented in Figure 16. In addition, ARiana was expected to save annotation time from the field workers. Therefore, in this study we defined two metrics, savings factor A and savings factor B (shown in Eqs.3 and 4) to study how quickly the user could use ARiana to complete annotation.

$$\text{savings factor A} = \frac{AjalonTime - ARianaTime}{ARianaTime} \tag{3}$$

where *AjalonTime* and *ARianaTime* represent the time spent annotating using Ajalon and ARiana, respectively. *ARianaTime* only counts the time when participants were creating annotations, such as giving voice commands, pointing to objects, and recording instructions. It includes the time when the user was providing voice input while conducting assembly tasks.

$$\text{savings factor B} = \frac{(task1 + AjalonTime) - (task2 + ARianaTime)}{(task2 + ARianaTime)} \tag{4}$$

where $task1$ represents the time spent for completing the assembly process that had been demonstrated and recorded by an expert. The recorded video was used for the Ajalon tool. $task2$ represents the time spent only for the same assembly process while using ARiana. The main reason that we defined savings factor B was to determine whether multitasking in ARiana slowed the entire process of collection and annotation.

From a user experience perspective, we measured ease-of-use and cognitive load. For the former, we used a custom questionnaire (5-point Likert scale) to evaluate the ease-of-use of ARiana in creating different types of annotations. For the latter, we selected NASA-TLX as a questionnaire and used it un-weighted with a 7-point Likert scale. More specifically, we explored the mental demand, physical demand, temporal demand, performance, effort, and frustration level of the participants during the test. We chose NASA-TLX because it has commonly been used for measuring both physical and cognitive workloads with tools [73], [74], [75].

3) DG3

ARiana provides low-latency responses through edge offloading. In addition to latency measurements, the satisfaction levels of users were collected to evaluate user' expectations for ARiana's latency of response.

B. STUDY DESIGN

The study was conducted in a laboratory environment simulating an assembly task in an industrial context using a wooden toy set (see Figure 10). The toy set was chosen for the following reasons:

First, it provided a representative task that could be generalised across use cases. ARiana was designed for capturing and annotation of manual maintenance and assembly

FIGURE 16. An example of annotation error when labeling objects association list. For Ajalon, (A) correct labeling (B) incorrect labeling due to occlusion. For ARiana, (C) correct labeling (D) incorrect labeling.

processes in general. Therefore, we chose a task that included both the use of hands and manipulation of graspable objects. The toy set was designed to simulate a real mechanical toolbox commonly used in industrial contexts. With nuts, bolts, screws, or the screwdriver shown in Figure 10, several industrial maintenance processes could be simulated easily. The toy set provided such and also allowed reflection of our results across use cases.

Second, the toy set allowed us to use novice participants to investigate how well ARiana could support new users in learning to annotate. Real maintenance tasks would have required field experts for those tasks. In addition, real maintenance tasks also contain physical risks. As a result, it did not suit this evaluation.

Third, the toy set allowed experimental control. For example, it allowed control of the length of the process, such as in evaluation of a 5-step process, a 10-step process, or even a 20-step process. Therefore, the task load measure was exposed to less bias than in a real assembly or maintenance case. In practical cases, the length of an assembly or maintenance process may vary due to continuous changes in the working environment, such as interactions other workers or with the system that is being maintained. The toy set was static and controllable for these aspects.

1) TASKS AND PARTICIPANTS

The study involved annotating videos of the same assembly process using either ARiana or Ajalon. The task with Ajalon in our study was to annotate a video pre-recorded by a field expert. The tasks with the two tools were conducted in separate sessions to avoid bias due to memory of the assembly process. There was a gap of at least 7 days between the two sessions. Each session lasted for approximately 45 minutes.

We recruited 18 participants, including six females, for the user study. All except one were graduate and post-graduate students from a local university. 94% of the participants had never used AR glasses or had used them only a few times. All of them had never used gestural interaction or had little experience with it. Before the study started, all the participants reviewed an informed consent form and gave their consent for participation. 13 participants conducted the tasks first with Ajalon and the remaining five first with ARiana.

2) PROCEDURES

Here we explain the procedures of video annotations using ARiana and Ajalon.

ARiana All participants were introduced to the role of a field expert. In that role, the participants were first asked to practice for 5 minutes to become familiar with the final product of the assembly tasks (see Figure 17f) and the five steps, as illustrated in Figure 17b-f, to build the product. Each participant attempted to build the same product but could freely change the assembly steps. They were told to use ARiana for recording and annotating the video with workflow information during the assembly process. After completing the assembly and annotation tasks, each participant completed the two questionnaires listed in Table 3. Finally, the participants engaged in a semi-structured interview to explore their experiences in creating the three types of annotations (defined in Section II.A) with ARiana.

Ajalon The participants were first introduced to the Ajalon user interface (see Figure 3) by explaining its main features for video annotation. For the task, the participants were introduced to a role of a video annotator, which they were to assume. They were then given a video that had captured the assembly process of a field expert and told that their task was to annotate it using Ajalon. The video had been uploaded to Ajalon's server beforehand and the participants could access Ajalon via a URL.

The participants received the following three sub-tasks to perform the entire process of video annotating: 1) splitting the prerecorded video of a toy model assembly into different working steps, 2) labeling all tools used, and 3) providing an instruction for each working step. Participants were recommended to follow the tasks in the given order. At the beginning of the task, they were encouraged to think aloud, which involved verbalizing their thought processes throughout the entire task. The purpose of the think-aloud method was to help us to understand their behavior during the annotation process.

Upon task completion, participants completed two questionnaires, as in the test of ARiana. After that step, the participants were interviewed briefly with open-ended questions to reflect on the difficulty of each of the three tasks and the problems they had encountered.

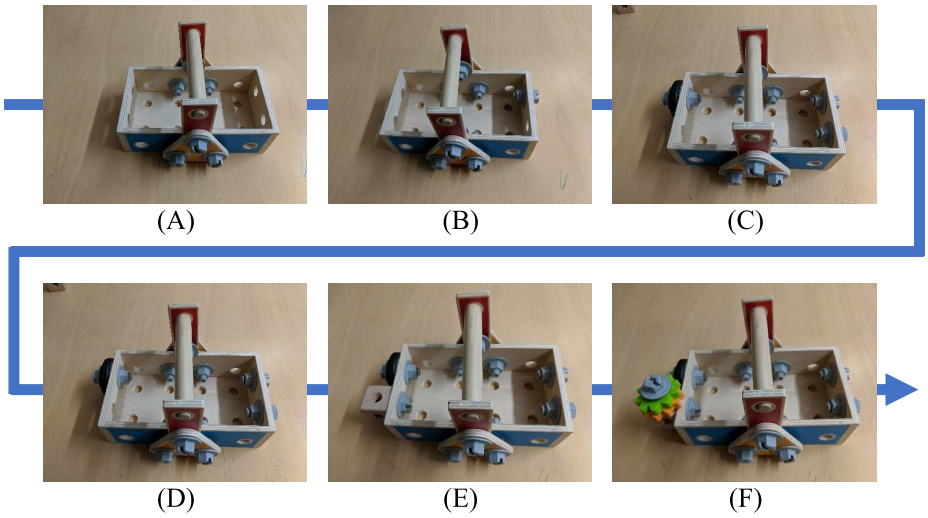


FIGURE 17. The assembly task for the participants. (A) The initial stage. (B-E) Intermediate stages.(F) The final stage.

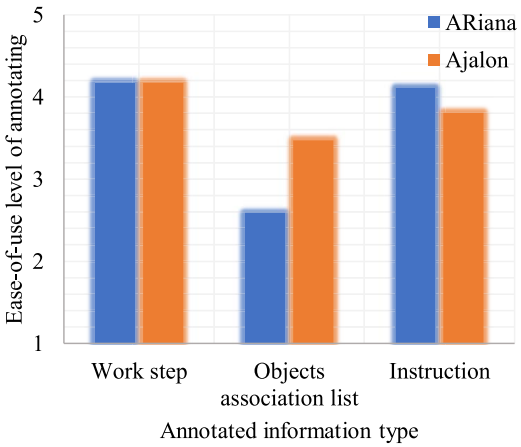


FIGURE 18. Ease-of-use: ARiana vs. Ajalon.

Due to restrictions on physical meetings during the COVID-19 pandemic, some participants conducted the study with Ajalon remotely via a video call. Those participants were asked to share their screen during the study, and with their consent, audio and video were recorded.

C. RESULTS OF USABILITY TESTS

1) EASE-OF-USE

According to the results of the customized questionnaire, ARiana is in general a user-friendly in-situ video annotation tool, with the overall score of 3.62 over 5.00. When comparing ARiana with Ajalon, as shown in Figure 18, ARiana felt easier to use for labeling work steps and giving instructions.

However, participants felt it easier to use Ajalon to annotate objects, including tools and materials used in each step. According to the feedback received from the participants, 11 experienced a long delay in labelling objects using gesture and speech. The delay was mainly caused by varying network latency of video streaming in a public network (25 - 562 ms) as well as errors in recognition of pointing gestures due to the invisibility of hands. Figure 19 shows example scenarios in which the users did not notice that their hand was not captured by the camera, which resulted in failure of gesture recognition and further failure of locating and labeling objects of interest in time.

2) EFFICIENCY

According to the time logs of the annotation sessions, the ratio of annotating time to the total length of the recorded video was 0.7, with a standard deviation of 0.15, when ARiana was in use. The time log also uncovered clear advantages of using ARiana over Ajalon. On average, participants spent 17.86 minutes to annotate the assembly process with Ajalon, compared with 5.2 minutes with ARiana. This time was measured as the total length of annotation activity of the participant in a session without the assembly time. In terms of savings factor A, as shown in Figure 20a, the value of savings factor A varied between 0.6 and 8.4. The variation between participants was large and correlated with the user's prior experience with AR and gesture and voice commands. For example, the three participants with the highest savings factor A, namely, P1, P5, and P16, were all experienced with AR headsets, gesture commands, and voice commands. In contrast, participants P4, P8, and P9 had experience with voice commands or AR headsets only. For example, P9 had used AR headsets a few times but had never used gesture commands or voice commands before. P4 and P8 had only experienced voice commands.

As shown in Figure 20b, when considering the total amount of time spent on recording and annotation, ARiana still saved time significantly. Savings factor B values ranged between 0.6 and 5.2. With assembly time, which is significantly

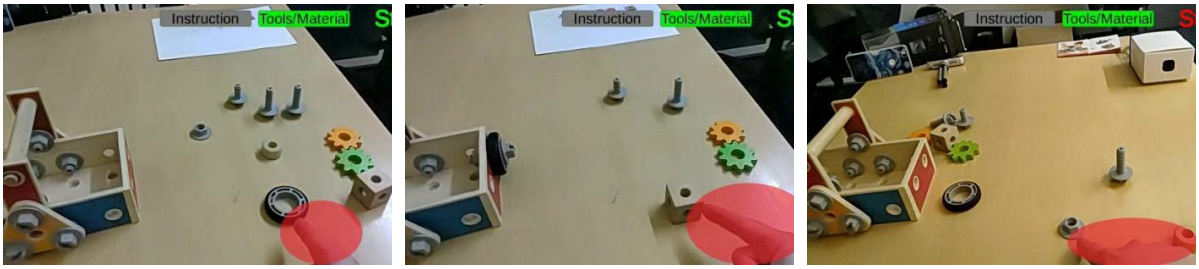


FIGURE 19. Hand-occlusion issues experienced by participants in the test of ARiana. The hand regions are marked with red color.

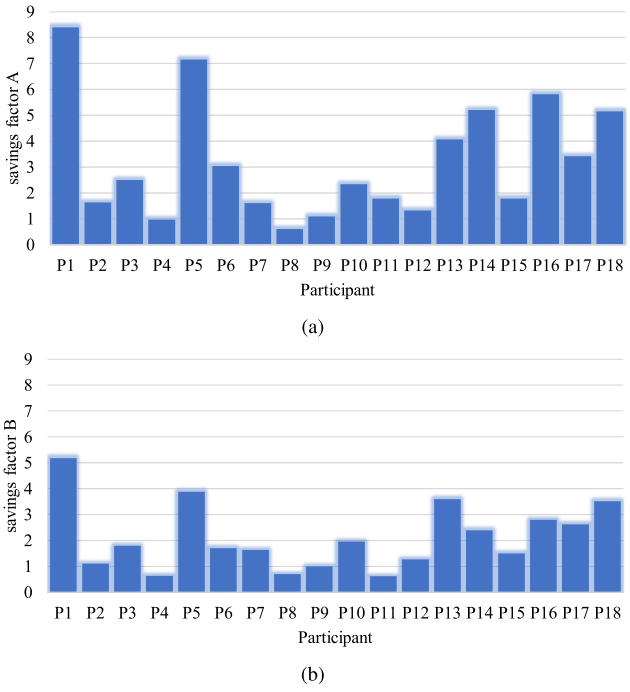


FIGURE 20. Savings factors of using ARiana over Ajalon with 18 participants. (a) Savings factor A. (b)Savings factor B.

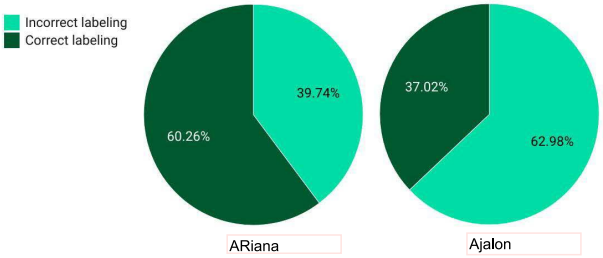


FIGURE 21. Annotation accuracy in the case the object association list: ARiana vs. Ajalon.

greater than annotation time, ARiana still has a better performance. So, ARiana supported significantly the whole process of demonstrating the assembly and annotating video.

The time savings was achieved in two ways. One way was to detect idle states accurately in order to minimize the idle time included in the collected video. For the video collected using ARiana, only 2.8% of the duration contained non-informative content (i.e., idle period in this case). The other was to propose annotations with the support of context awareness in order to reduce the time spent creating annotations. The benefit of in-situ annotation was stated by participant P6, who stated that they did not need to re-watch the demonstration video and recall the process; rather, they only needed to demonstrate the task and annotate the workflow at the same time. Regarding the process of understanding, participant P4 mentioned that the annotation was simple, and they were able to annotate as naturally as they instructed other persons to complete the assembly process.

From these results, we claim that ARiana is more efficient than Ajalon in terms of annotation time.

3) EFFECTIVENESS

Annotation accuracy was used to measure the effectiveness of ARiana. Because work-step identification and the instruction for each work step can be subjective, we decided to compare only the quantitative measurement of errors in object annotation (i.e., the creation of the object association list) between ARiana and Ajalon.

As illustrated in Figure 21, users tended to make fewer mistakes when editing the object association list using gestures and speech in the case of ARiana than mouse and keyboard in the case of Ajalon. It is noteworthy that here we only compare the annotations that were edited manually. Regarding the number of manually created annotations, participants created 234 using ARiana and 235 using Ajalon.

The difference in terms of the annotation quality was partially affected by the number of occlusions appearing in the videos. Out of the 234 manual annotations collected with ARiana, only 4.7% of the annotated objects had been partially occluded by other objects. However, in separating video recording from annotation, the objects had often been occluded by hands or other parts of the body in the video, as the field expert could not always pay attention to occlusions.

4) COGNITIVE LOAD

Figure 22 compares the cognitive load between ARiana and Ajalon based on the results of the NASA-TLX test. In general, there was no significant difference between these two cases.

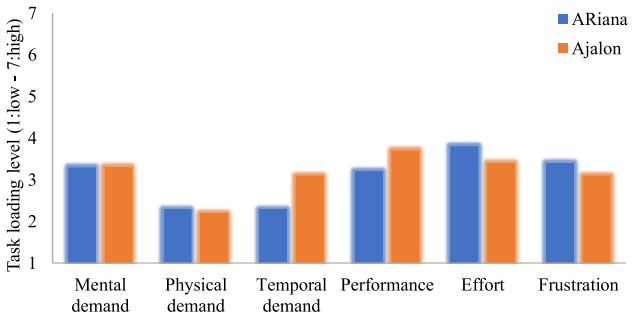


FIGURE 22. NASA-TLX scores for mental demand, physical demand, temporal demand, performance, effort, and frustration when using ARiana or Ajalon.

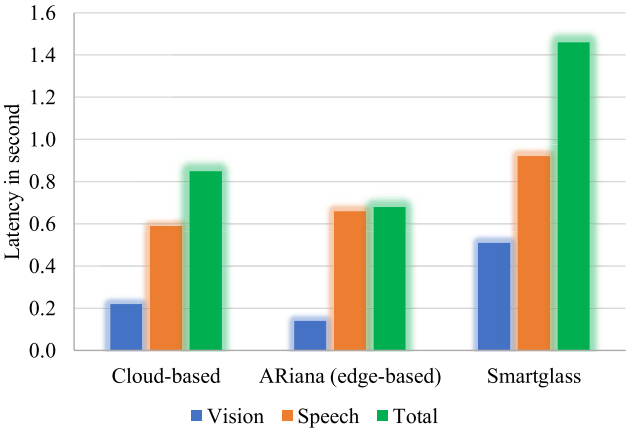


FIGURE 23. Comparison of latencies among different setups. They are end-to-end latency including also networking latency in audio processing, video processing and total.

Thus, ARiana appears to be able to maintain the cognitive load in a two-task scenario as low as that of a one-task scenario in the case of Ajalon.

The context-awareness module of ARiana contributed partially to reducing the cognitive load of participants. The qualitative findings revealed that approximately 13% of annotations had been proposed by the context-awareness module of ARiana. Seven out of 18 participants mentioned that ARiana reminded them to provide annotations. This made participants feel less anxious in remembering to produce annotation while demonstrating. Participants P5 and P6 also emphasized that it was easier for them to annotate information by answering the prompts of ARiana. Moreover, it took less time for them to provide annotation.

In addition, the elicited multimodal interface is another essential piece of ARiana that kept the mental and physical demand as that of Ajalon even though users had to carry out both demonstration and annotation of the task simultaneously. Ten out of 18 participants said that it felt comfortable to provide annotations for each step segment, object association list, and instruction at the time of demonstrating the task. In addition, they also mentioned that the visual feedback of the interface, including panels on the top and object bounding box, allowed them to continue tracking the

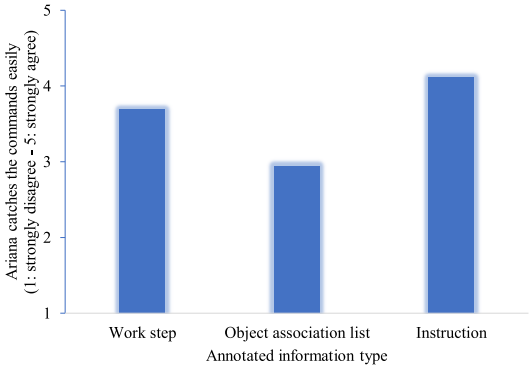


FIGURE 24. User satisfaction along the latency of ARiana for each annotated information type.

annotation progress easily while demonstrating the assembly task. Moreover, four out of 18 participants emphasized they were confident in performing both demonstration and annotation at the same time with the designed multimodal interaction.

5) LATENCY

Figure 23 compares the response delay of in-situ annotation among three setups, including local processing on glasses, computation offloading to the edge (implemented in ARiana), and computation offloading to the cloud. The deployed hardware included an edge device (CPU: Intel Core I7-7820HK, GPU: NVIDIA GeForce GTX1070, RAM: 16GB), and a powerful cloud computer (CPU: Intel Xeon W-2133, GPU: Nvidia Quadro P5000, RAM: 64GB). ARiana consists of vision-related and speech-related processing. The vision-related component includes hand-object interaction detection and pointing-gesture recognition, while the speech-related component includes speech-to-text object-name extraction for labeling and recognition of voice commands. For local processing, both are run on smart glasses. In the edge-based setup, vision-related processing is performed at the edge device, while speech-related processing is offloaded to the IBM Watson cloud service [76]. In the cloud setup, both vision-related and speech-related components are implemented in the cloud. As a result, the edge-based deployment performed significantly better than the others, with an average end-to-end latency of 660 ms as shown in Figure 23. That level of latency is sufficient to assist users in completing state-flow tasks such as *sandwich making* and *Lego assembly* [70] because the expected latency for those applications is in the range of $600 - 2700$ ms by user evaluation. However, Figure 24 shows that ARiana only achieved average satisfaction of users, which is below the average with 2.94, in annotating associated objects as compared with work steps and instructions. Over 70% of participants answered that they felt unsatisfied by the delay in annotating associated objects using pointing gestures and voice.

Based on latency logs, the network transmission of the vision-related component caused the latency of approximately 78 ms while the processing latency on the edge was

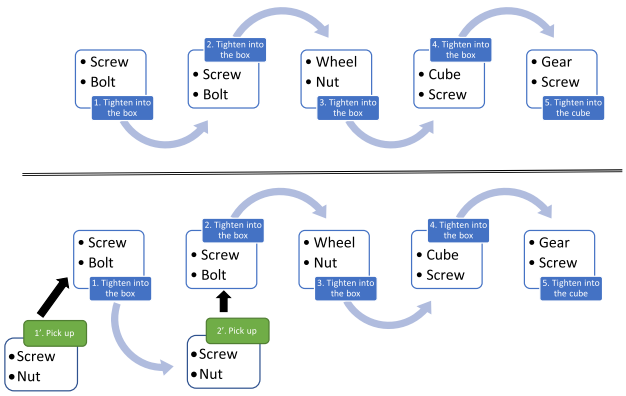


FIGURE 25. Comparison of workflow taken by two different users to complete the same assembly task described in Figure 17. The green boxes represent the additional steps the user under the line added to the workflow.

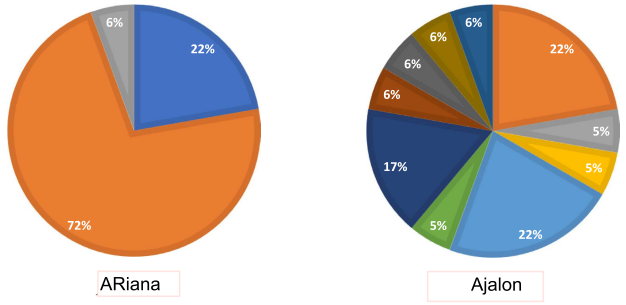


FIGURE 26. Variation in workflow: ARiana versus Ajalon. Each color represents a variant.

about 62 ms. Based on these observations, the cloud-based speech recognition appears to be a key obstacle. One way to reduce the latency would be to deploy the speech recognition platform on the edge as well to shorten the network delay. Moreover, with upcoming 5G and beyond networks, it should be possible to reduce the wireless communication delays to further reduce network latency.

According to the summary of the results in Table 3, ARiana has proven to be an easy-to-use in-situ video annotation tool that guides users to create high-quality annotations in an efficient way. The design of ARiana has fulfilled the three DGs and has shown advantages over Ajalon. During the experiment, we have made the following observations about the variation in workload and interactivity of ARiana as opposed to Ajalon.

6) VARIATION IN WORKFLOW

Participants were supposed to complete the same task. In practice, the way they executed the task varied. Thus, they may have planned the workflow differently in terms of the number of steps and the order of steps. For example, Figure 25 compares the workflow taken by two different users to complete the same assembly task, and shows that one user defined two additional steps.

We compared the execution among participants, and analyzed the variation in workflow when using ARiana or Ajalon. As shown in Figure 26, 11 variants were identified from the test results of Ajalon. In the case of ARiana, the number of variants converged to three with a dominant one shared among 14 participants. After analyzing the dominant variant, we realized that the convergence occurred because participants had understood deeply the demonstration when using ARiana. This convergence helps novices gain a deep understanding of assembly and maintenance processes after watching annotated videos.

VII. SUMMARY

In the evaluation of ARiana, we found that it improved performance and user experience in video annotation over a state-of-the-art tool in many aspects. Here, we summarize the evaluation results and the implications of these results in the design of novel video annotation tools.

Firstly, with Ariana, labeling is now integrated with demonstration of a task. This means the user can adjust the camera setup and move the objects to annotate information easily in order to obtain a high-quality video. With an unoccluded first-person view, the machine is more reliable and efficient in object annotation than humans are with third-person-perspective recording.

Secondly, ARiana saves annotation time by automatically removing non-informative parts from recorded video and by proposing annotations of objects and work-step segmentation to reduce the manual efforts needed for adding annotations while maintaining the cognitive load moderate as Ajalon.

Thirdly, compared with Ajalon, ARiana improved the ease-of-use in two out of 3 types of annotation tasks due to its multimodal interaction design.

Fourthly, ARiana keeps latency low enough to achieve useful guidance for annotation.

To further refine the design goals, as well as to gain other important insights for improving ARiana, we recommend the following next steps:

- Implementing speech recognition at the edge to reduce the latency significantly;
- Detecting whether hands have come out of the viewpoint of the first-person camera view in real time and guiding the user to fix the problem by adjusting the camera pose or the body position;
- Improving the efficiency of the computer vision algorithms involved to reduce the response delay;
- Testing ARiana in real-world industrial environments to validate the findings from this laboratory study.

VIII. CONCLUSION

To address problems in the conventional process of recording and annotating assembly videos, we have developed ARiana, a novel in-situ video annotation tool running on smart glasses. ARiana allows field experts to efficiently annotate first-person videos with workflow information (i.e., starting and ending points of all work steps, tools and materials used, and instructions for each step) while conducting assembly or maintenance tasks. By conducting a user study with

18 participants using a toy set model, ARiana proved to be more efficient and effective than a state-of-the-art video annotation tool. In particular, with context awareness enabled, ARiana is able to guide users to create annotations with less effort and higher accuracy. Furthermore, through the combination of voice and gestural interactions, ARiana provides a natural way for users to complete annotation tasks without causing excessive cognitive load. However, ARiana still needs to be evaluated more in several industrial environments to show its usability across several contexts. In addtion, latency requires to be studied and improved for a better user experience when using ARiana.

APPENDIX A QUESTIONS IN FORMATIVE INTERVIEW

- What are the difficult problems you commonly have while annotating or documenting a maintenance video?
- How do you feel about those difficult problems?
- Have you ever thought about the solution for those problems? What are those solutions?
- How can you setup cameras for recording maintenance-task demonstration?
- How do you feel when annotating a long video?
- How do you feel about the impact of those videos on annotating or documenting process?

APPENDIX B QUESTIONNAIRE FOR ELICITATION STUDY IN INTERACTION DESIGN OF ARiana

A. VOICE COMMANDS

1) It is easy to use selected voice command for ''Start annotating'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
2) It is easy to memorize selected voice command for ''Start annotating'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
3) It is easy to use selected voice command for ''Stop annotating'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
4) It is easy to memorize selected voice command for ''Stop annotating'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
5) It is easy to use selected voice command for ''Start a new step'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
6) It is easy to memorize selected voice command for ''Start a new step'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
7) It is easy to use selected voice command for ''Start to give instruction'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
8) It is easy to memorize selected voice command for ''Start to give instruction'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
9) It is easy to use selected voice command for ''Undo the latest annotating event'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
10) It is easy to memorize selected voice command for ''Undo the latest annotating event'' command.
    (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree

B. GESTURE

1) It is easy to use selected gesture for ''Start annotating'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
2) It is easy to memorize selected gesture for ''Start annotating'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
3) It is easy to use selected gesture for ''Stop annotating'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
4) It is easy to memorize selected gesture for ''Stop annotating'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
5) It is easy to use selected gesture for ''Start a new step'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
6) It is easy to memorize selected gesture for ''Start a new step'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
7) It is easy to use selected gesture for ''Present an object'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
8) It is easy to memorize selected gesture for ''Present an object'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
9) It is easy to use selected gesture for ''Start to give instruction'' command.
   (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
10) It is easy to memorize selected gesture for ''Start to give instruction'' command.
    (a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
11) It is easy to use selected gesture for ''Undo the latest annotating event'' command.

(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree

12) It is easy to memorize selected gesture for "Undo the latest annotating event" command.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree

APPENDIX C QUESTIONNAIRE FOR THE EVALUATING EASE-OF-USE OF AJALON AND ARiana

C. AJALON

1) It is easy to annotate work step with Ajalon.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
2) It is easy to annotate objects association list with Ajalon.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
3) It is easy to annotate instruction with Ajalon.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree

D. ARiana

1) It is easy to annotate work step with ARiana.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
2) It is easy to annotate objects association list with ARiana.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
3) It is easy to annotate instruction with ARiana.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree

APPENDIX D QUESTIONNAIRE FOR EVALUATING THE SATISFACTION OF USERS FOR THE ARiana USAGE

1) ARiana easily catched your work step annotating command.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
2) ARiana easily catched your object annotating commands.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
3) ARiana easily catched your instruction commands.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree
4) You felt satisfied with the latency of the visual feedback while using ARiana.
(a) strongly disagree (b) disagree (c) neutral (d) agree (e) strongly agree

REFERENCES

[1] E. McQuade, E. Sjoer, P. Fabian, J. Carlos Nascimento, and S. Schroeder, "Will you miss me when I'm gone?" *J. Eur. Ind. Training*, vol. 31, no. 9, pp. 758–768, Nov. 2007.
[2] D. Leonard, W. Swap, and G. Barton, "What's lost when experts retire," Harvard Business Review, Dec. 2014.
[3] B. A. Tunanidis, "Baby boomers retirement: Succession planning for the next generation of legacy systems projects," Ph.D. thesis, Northcentral Univ., San Diego, CA, USA, 2018.
[4] A. D. P. dos Santos, L. Loke, and R. Martinez-Maldonado, "Exploring video annotation as a tool to support dance teaching," in *Proc. 30th Austral. Conf. Comput.-Hum. Interact.*, Dec. 2018, pp. 448–452.
[5] K. C. Leung and M. P. Shek, "Adoption of video annotation tool in enhancing students' reflective ability level and communication competence," *Coaching: Int. J. Theory, Res. Pract.*, vol. 14, no. 2, pp. 151–161, 2021.
[6] X. Ke, J. Zou, and Y. Niu, "End-to-end automatic image annotation based on deep CNN and multi-label data augmentation," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2093–2106, Aug. 2019.
[7] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognit.*, vol. 79, pp. 242–259, Jul. 2018.
[8] S. S. Aote and A. Potnurwar, "An automatic video annotation framework based on two level keyframe extraction mechanism," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 14465–14484, Jun. 2019.
[9] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vis.*, vol. 101, no. 1, pp. 184–204, Jan. 2013.
[10] S. Anjum, C. Lin, and D. Gurari, "CrowdMOT: Crowdsourcing strategies for tracking multiple objects in videos," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, pp. 1–25, Jan. 2021.
[11] J. Yuen, B. Russell, C. Liu, and A. Torralba, "LabelMe video: Building a video database with human annotations," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1451–1458.
[12] D. Melhart, A. Liapis, and G. N. Yannakakis, "PAGAN: Video affect annotation made easy," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 130–136.
[13] M. N. Amorim, F. R. A. Neto, and C. A. S. Santos, "Achieving complex media annotation through collective wisdom and effort from the crowd," in *Proc. 25th Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jun. 2018, pp. 1–5.
[14] T. A. Pham, J. Wang, R. Iyengar, Y. Xiao, P. Pillai, R. Klatzky, and M. Satyanarayanan, "Ajalon: Simplifying the authoring of wearable cognitive assistants," *Softw., Pract. Exper.*, vol. 51, no. 8, pp. 1773–1797, May 2021.
[15] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4575–4583.
[16] *Furniture Assembly Miami*. Accessed: May 5, 2021. [Online]. Available: https://furnitureassemblypros.com/furniture-assembly-miami
[17] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "COIN: A large-scale dataset for comprehensive instructional video analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1207–1216.
[18] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3537–3545.
[19] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, and D. Kalinin, "opencv/cvat: V1.1.0," 2020. Accessed: Sep. 1, 2021, doi: 10.5281/zenodo.4009388.
[20] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2276–2279.
[21] *Labelbox*. Accessed: Dec. 12, 2020. [Online]. Available: https://labelbox.com/
[22] *Pixel Annotation Tool*. Accessed: Dec. 12, 2020. [Online]. Available: https://github.com/abreheret/PixelAnnotationTool
[23] F. Ragusa, A. Furnari, S. Livatino, and G. M. Farinella, "The MECCANO dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1569–1578.
[24] G. A. Lee, S. Ahn, W. Hoff, and M. Billinghurst, "Enhancing first-person view task instruction videos with augmented reality cues," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Nov. 2020, pp. 498–508.
[25] D. Damen, T. Leelasawassuk, and W. Mayol-Cuevas, "You-do, I-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance," *Comput. Vis. Image Understand.*, vol. 149, pp. 98–112, Aug. 2016.
[26] J. Zubizarreta, I. Aguinaga, and A. Amundarain, "A framework for augmented reality guidance in industry," *Int. J. Adv. Manuf. Technol.*, vol. 102, nos. 9–12, pp. 4095–4108, Jun. 2019.

[27] F. Bruno, L. Barbieri, E. Marino, M. Muzzupappa, L. D'Oriano, and B. Colacino, "An augmented reality tool to detect and annotate design variations in an industry 4.0 approach," *Int. J. Adv. Manuf. Technol.*, vol. 105, nos. 1–4, pp. 875–887, Nov. 2019.
[28] S. Guven and S. Feiner, "Authoring 3D hypermedia for wearable augmented and virtual reality," in *Proc. 7th IEEE Int. Symp. Wearable Comput.*, Oct. 2003, pp. 21–23.
[29] Y. Yaginuma, M. Furukawa, and T. Yamada, "Video annotation tool for learning job interview," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, Mar. 2017, pp. 534–535.
[30] P. Rennert, O. M. Aodha, M. Piper, and G. Brostow, "VideoTagger: User-friendly software for annotating video experiments of any duration," *bioRxiv*, pp. 272–468, Mar. 2018.
[31] A. Nassani, H. Kim, G. Lee, M. Billinghurst, T. Langlotz, and R. W. Lindeman, "Augmented reality annotation for social video sharing," in *Proc. SIGGRAPH ASIA Mobile Graph. Interact. Appl.*, 2016, pp. 1–5.
[32] M. Gygli and V. Ferrari, "Efficient object annotation via speaking and pointing," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1061–1075, May 2020.
[33] C. Auepanwiriyakul, A. Harston, P. Orlov, A. Shafti, and A. A. Faisal, "Semantic fovea: Real-time annotation of ego-centric videos with gaze context," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, 2018, pp. 1–3.
[34] H. Bai, P. Sasikumar, J. Yang, and M. Billinghurst, "A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–13.
[35] A. Berg, J. Johnander, F. Durand de Gevigney, J. Ahlberg, and M. Felberg, "Semi-automatic annotation of objects in visual-thermal video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2242–2251.
[36] I. Wang, P. Narayana, J. Smith, B. Draper, R. Beveridge, and J. Ruiz, "Easel: Easy automatic segmentation event labeler," in *Proc. 23rd Int. Conf. Intell. User Interfaces*, 2018, pp. 595–599.
[37] L.-H. Lee and P. Hui, "Interaction methods for smart glasses: A survey," *IEEE Access*, vol. 6, pp. 28712–28732, 2018.
[38] Y.-C. Tung, C.-Y. Hsu, H.-Y. Wang, S. Chyou, J.-W. Lin, P.-J. Wu, A. Valstar, and M. Y. Chen, "User-defined game input for smart glasses in public space," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 3327–3336.
[39] Y. Li, Y. Cheng, W. Meng, Y. Li, and R. H. Deng, "Designing leakage-resilient password entry on head-mounted smart wearable glass devices," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 307–321, 2021.
[40] J. Lee, H.-S. Yeo, M. Dhuliawala, J. Akano, J. Shimizu, T. Starner, A. Quigley, W. Woo, and K. Kunze, "Itchy nose: Discreet gesture interaction using EOG sensors in smart eyewear," in *Proc. ACM Int. Symp. Wearable Comput.*, 2017, pp. 94–97.
[41] C. J. Liu, C.-L. Yang, J. J. Williams, and H.-C. Wang, "Notestruct: Scaffolding note-taking while learning from online videos," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, 2019, p. 1116.
[42] M. N. D. Amorim, R. M. Segundo, C. A. Santos, and O. D. L. Tavares, "Video annotation by cascading microtasks: A crowdsourcing approach," in *Proc. 23rd Brazillian Symp. Multimedia Web*, 2017, pp. 49–56.
[43] A. Shen, "Beaverdam: Video annotation tool for computer vision training labels," Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2016-193, 2016.
[44] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "The EPIC-KITCHENS dataset: Collection, challenges and baselines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4125–4141, Nov. 2021.
[45] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.
[46] T. M. Ward, D. M. Fer, Y. Ban, G. Rosman, O. R. Meireles, and D. A. Hashimoto, "Challenges in surgical video annotation," *Comput. Assist. Surg.*, vol. 26, no. 1, pp. 58–68, Jan. 2021.
[47] Y. Huang, M. Cai, H. Kera, R. Yonetani, K. Higuchi, and Y. Sato, "Temporal localization and spatial segmentation of joint attention in multiple first-person videos," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2313–2321.
[48] Y.-C. Su and K. Grauman, "Detecting engagement in egocentric video," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Copenhagen, Denmark: Springer, 2016, pp. 454–471.
[49] F. Riillo, L. R. Quitadamo, F. Cavrini, G. Saggio, C. A. Pinto, N. C. Pasto, L. Sbernini, and E. Gruppioni, "Evaluating the influence of subject-related variables on EMG-based hand gesture classification," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2014, pp. 1–5.
[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[51] K. Lee and H. Kacorri, "Hands holding clues for object recognition in teachable machines," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–12.
[52] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe hands: On-device real-time hand tracking," 2020, *arXiv:2006.10214*.
[53] X. Huang and Y.-J. Zhang, "300-FPS salient object detection via minimum directional contrast," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4243–4254, Sep. 2017.
[54] J. A. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 702–715.
[55] E. Chan, T. Seyed, W. Stuerzlinger, X.-D. Yang, and F. Maurer, "User elicitation on single-hand microgestures," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 3403–3414.
[56] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn, "User-defined gestures for augmented reality," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst. (CHI EA)*, 2013, pp. 282–299.
[57] H. Dong, N. Figueroa, and A. El Saddik, "An elicitation study on gesture attitudes and preferences towards an interactive hand-gesture vocabulary," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 999–1002.
[58] L. S. Figueiredo, M. G. M. Gonçalves Maciel Pinheiro, E. X. C. Vilar Neto, and V. Teichrieb, "An open catalog of hand gestures from sci-fi movies," in *Proc. 33rd Annu. ACM Conf. Extended Abstr. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 1319–1324.
[59] F. Kerber, M. Puhl, and A. Krüger, "User-independent real-time hand gesture recognition based on surface electromyography," in *Proc. 19th Int. Conf. Hum.-Comput. Interact. with Mobile Devices Services*, Sep. 2017, pp. 1–7.
[60] A. Alanwar, M. Alzantot, B.-J. Ho, P. Martin, and M. Srivastava, "Selecon: Scalable IoT device selection and control using hand gestures," in *Proc. 2nd Int. Conf. Internet–Things Design Implement.*, 2017, pp. 47–58.
[61] Y. Zhang, Y. Chen, H. Yu, X. Yang, W. Lu, and H. Liu, "Wearing-independent hand gesture recognition method based on EMG armband," *Pers. Ubiquitous Comput.*, vol. 22, no. 3, pp. 511–524, Jun. 2018.
[62] B. Walther-Franks, T. Döring, M. Yilmaz, and R. Malaka, "Embodiment or manipulation? Understanding users' strategies for free-hand character control," in *Proc. Mensch und Comput.*, 2019, pp. 661–665.
[63] D. Park, Y. S. Lee, M. H. Yun, S. Song, I. Rhiu, S. Kwon, and Y. An, "User centered gesture development for smart lighting," in *Proc. HCI Korea*, Jan. 2016, pp. 146–150.
[64] A. Boudjelthia, S. Nasim, J. Eskola, J. M. Adeegbe, O. Hourula, S. Klakegg, and D. Ferreira, "Enabling mid-air browser interaction with leap motion," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 335–338.
[65] M. Chang, A. Truong, O. Wang, M. Agrawala, and J. Kim, "How to design voice based navigation for how-to videos," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–11.
[66] R.-D. Vatavu and J. O. Wobbrock, "Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 1325–1334.
[67] *IBM Watson Cloud Service*. Accessed: Dec. 12, 2020. [Online]. Available: https://cloud.ibm.com/catalog/services/speech-to-text
[68] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*.
[69] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *Proc. 3rd Workshop Comput. Vis. AR/VR IEEE Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1–4.
[70] Z. Chen, W. Hu, J. Wang, S. Zhao, B. Amos, G. Wu, K. Ha, K. Elgazzar, P. Pillai, R. Klatzky, D. Siewiorek, and M. Satyanarayanan, "An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance," in *Proc. 2nd ACM/IEEE Symp. Edge Comput.*, Oct. 2017, pp. 1–14.
[71] L. Liu, H. Li, and M. Gruteser, "Edge assisted real-time object detection for mobile augmented reality," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, Aug. 2019, pp. 1–16.

[72] E. D. W.-D. Vries, W. A. Dolfsma, H. J. V. D. Windt, and M. P. Gerkema, "Knowledge transfer in university–industry research partnerships: A review," *J. Technol. Transf.*, vol. 44, no. 4, pp. 1236–1255, 2019.
[73] K. Ikematsu and S. Yamanaka, "ScraTouch: Extending interaction technique using fingernail on unmodified capacitive touch surfaces," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 3, pp. 1–19, 2020.
[74] H. Ye, K. C. Kwan, W. Su, and H. Fu, "ARAnimator: In-situ character animation in mobile AR with user-defined motion gestures," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 1–83, 2020.
[75] Y.-P. Yau, L. H. Lee, Z. Li, T. Braud, Y.-H. Ho, and P. Hui, "How subtle can it get? A trimodal study of ring-sized interfaces for one-handed drone control," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 2, pp. 1–29, 2020.
[76] (2020). *IBM Watson Cloud Service*. Accessed: Dec. 12, 2022. [Online]. Available: https://cloud.ibm.com/catalog/services/speech-to-text

TRUONG AN PHAM received the B.Sc. degree in computer science from the University of Science, Vietnam, in 2013, and the M.E. degree in multimedia technology from Telecom ParisTech, France, in 2016. He is currently pursuing the Ph.D. degree in electrical engineering with Aalto University. His research interests include human–computer interaction, computer vision, and machine learning.

TIM MOESGEN received the M.A. degree in collaborative and industrial design from Aalto University, in 2022. He is currently pursuing the doctoral degree with the School of Electrical Engineering, Aalto University. His research interests include haptic interfaces, inclusive design, and wearable technology.

SANNI SILTANEN received the master's degree in applied mathematics from the University of Eastern Finland, in 1999, and the doctoral degree in information science from Aalto University, in 2015. She is the Senior Ecosystem Leader with Dimecc Ltd. and an Associate Professor with Tampere University. Her research interests include industrial extended reality (XR), mobile robotics, artificial intelligence, smart spaces, and human–technology interaction.

JOANNA BERGSTRÖM is an Associate Professor with the Department of Computer Science, University of Copenhagen. She develops interaction techniques, evaluation methods, and models to improve how people move with body-based user interfaces and in virtual reality. Her research interest includes human–computer interaction.

YU XIAO received the doctoral degree in computer science from Aalto University, in 2012. She is an Associate Professor with the School of Electrical Engineering, Aalto University. Her current research interests include edge computing, wearable sensing, and extended reality.