AR-Mentor: Augmented Reality Based Mentoring System

Zhiwei Zhu * Vlad Branzoi Michael Wolverton Glen Murray Nicholas Vitovitch Louise Yarnall
Girish Acharya Supun Samarasekera Rakesh Kumar

SRI International, USA

ABSTRACT

AR-Mentor is a wearable real time Augmented Reality (AR) mentoring system that is configured to assist in maintenance and repair tasks of complex machinery, such as vehicles, appliances, and industrial machinery. The system combines a wearable Optical-See-Through (OST) display device with high precision 6-Degree-Of-Freedom (DOF) pose tracking and a virtual personal assistant (VPA) with natural language, verbal conversational interaction, providing guidance to the user in the form of visual, audio and locational cues. The system is designed to be heads-up and hands-free allowing the user to freely move about the maintenance or training environment and receive globally aligned and context aware visual and audio instructions (animations, symbolic icons, text, multimedia content, speech). The user can interact with the system, ask questions and get clarifications and specific guidance for the task at hand. A pilot application with AR-Mentor was successfully built to instruct a novice to perform an advanced 33-step maintenance task on a training vehicle. The initial live training tests demonstrate that AR-Mentor is able to help and serve as an assistant to an instructor, freeing him/her to cover more students and to focus on higher-order teaching.

Index Terms: Mentoring System, Wearable Technology, Virtual Personal Assistant, Optical See Through Glasses

1 INTRODUCTION

1.1 Motivation

Recent research suggests that AR and VPA technologies offer the potential to achieve improved learning and on-site human-instructor-like guidance. When doing complex tasks, having in context visual and verbal guidance through the task allows faster and more efficient repair. This can be used both for training and in-the-field operations. AR enables the overlay of real-time visual information on a users view of the physical world, to guide him in performing tasks; VPA provides human-like interaction using spoken natural language, recognizes the users goals and provides feedback to the user in real-time. Together, these technologies are able to provide heads-up, hands-free operation that allows step-by-step, just-in-time guidance without being distracted by having to look at a technical manual or computer screen.

Based on these principles, we have developed and tested a prototype technology, named AR-Mentor. As shown in Figure 1, it is a head-mounted OST display device through which technicians can receive just-in-time audio-visual guidance while learning to conduct adjustments/repairs on equipment. AR-Mentor provides guidance via on-demand voice instruction and five types of visual overlays to the physical environment: 3D graphic animations to describe tools and components and how to manipulate them, animated arrows that direct the learner's gaze direction, live-action videos of maintainers conducting adjustment procedures, text-annotated graphic images of complex objects, and diagrammatic images of complex equipment

*e-mail: zhiwei.zhu, vlad.branzoi, michael.wolverton@sri.com

and vehicle map. The learner can interact with the system, ask questions and direct the flow and level of guidance, skipping known procedures and asking the system to repeat as needed. The device additionally features a capacity to detect when the learner is positioned correctly near/at the equipment to conduct an adjustment and provides feedback on correct placement.

Figure 1: Concept of the proposed AR-Mentor system: The user communicates verbally to the AR-Mentor system using a Microphone; The AR-Mentor system understands the user and provides audible (Speaker) and visual instructions (OST Glasses).

1.2 Related Work

Various AR-based training systems [5, 9, 8, 13, 15] have been proposed in the past years. AR for maintenance or assembly tasks [7, 17] has a long history, dating back to the original Boeing wire harness application [5]. However, as pointed out by Lee [9], there are still relatively few studies that have been done for the adoption and the usability of AR systems and innovations in industrial training.

Kim et al. [8] presented an AR e-training mobile application using smart phones and tablet PCs for car maintenance, and multimedia guidance information is displayed on the image of a real engine. Henderson et al. [7] presented a prototype AR application to support military mechanics conducting routine maintenance tasks inside an armored vehicle turret. Custom-built stereo VST HMD is used instead of OST HMD, and a wrist-worn controller is used to control the animations or cue the next task, which is not hands-free in a typically dirty environment. In addition, since they were focusing on the research instead of practical deployment, they installed 10

tracking cameras around the turret and installed 3 IR LEDs on the HMD to track the user's head movement. Webel et al. [17] developed a platform for multimodal AR-based training of maintenance and assembly skills to accelerate the technician's acquisition of new maintenance procedures. Instead of using an HMD, a mobile tablet equipped with a video camera is used for displaying visual instructions and vibrotactile bracelets are worn on the wrist to provide haptic feedbacks. Platonov et al. [15] developed a prototype AR system for industrial repair and maintenance. The system uses a markerless CAD based tracking method to display AR guidance and instructions on an HMD. A camera rigidly attached to the HMD provides the tracking. While the tracking uses pre-recorded real video keyframes it also relies on having a non-trivial CAD model in order to associate 2D to 3D points.

Compared to the existing mentoring systems, we combine the power of AR and VPA to create a heads-up, hands-free system very close to utilizing a human trainer with the best of the live and virtual training worlds. It has a host of capabilities that include: both vision and speech understanding for interacting with the user and his environment; has stored knowledge of a broad range of equipment used by the user; features a general reasoning capability to understand the user's objectives and what the user knows and needs to learn to reach those objectives and a sophisticated grasp of training techniques most likely to be effective with the user. This system not only ensures consistency in training, but also is available anywhere including in-theater.

The main goal of this paper is to present the system design and algorithms of an OST-AR and VPA based mentoring system. We have successfully developed an AR prototype system that can be used by student mechanics to learn and perform a 33-step vehicle maintenance task without any technical manuals or instructor. It can be configured easily to assist any other maintenance and repair tasks of vehicles, or any other complex machinery.

2 SYSTEM OVERVIEW

2.1 Hardware Setup

Figure 2 shows our own customized human wearable helmet-based sensor head package. Our sensor head package consists of one pair of stereo-cameras (Ximea xiQ MQ013MG-E2), one Inertial Measurement Unit (IMU) (Microstrain 3DM-GX3-25) and one Head-Mounted Monocular Display (Cyber-I SXGA Monocular HMD¹). The stereo cameras are arranged vertically for minimal intrusion to the user. The images (640x480) are captured at 15fps and the IMU unit operates at 100HZ. Basically, the stereo cameras and the IMU form a multi-sensor navigation unit to provide precise pose estimation. The cameras and the HMD are rigidly mounted together, and their spatial relationship can be calibrated in advance. Once the calibration is done, the pose estimated by the navigation unit can be transformed by the system to know where to insert the synthetic objects in the HMD. The current sensor-rig weights around 1 lb.

2.2 System Architecture

Figure 3 shows the key subsystems of the AR-Mentor system. Sensor Processing Subsystem (SPS) interfaces with the user worn sensors. This includes a microphone to capture speech and video/IMU based sensors to track the user's position, orientation and actions. The SPS block processes all the high-bandwidth, low-latency data to produce higher level information that are consumed by the down-stream subsystems. The audio feed is converted to textual phrases. The video feed along with the IMU data is interpreted to find the user's position with respect to the equipment and his gaze direction. The system also supports add-on modules for higher level constructs such as action recognition and object recognition.

¹<http://www.cybermindnl.com/products/cyber-i-series/cyber-i-sxga-monocular-hmd>

Figure 2: The human wearable helmet-based sensor package with OST display.

Figure 3: AR-Mentor system block diagram.

SPS coordinates the interactions with the other two sub-systems: Rendering Subsystem and the VPA subsystem. The VPA subsystem ingests the higher level constructs from the SPS to construct user intent. The intent is further analyzed using a knowledge-base that represents the task workflow to generate interactive content generated by AR-Mentor. VPA can also provide feedback to the SPS block on locales and actions of interest based on context. The VPA subsystem is setup as a stand-alone server that can be run remotely through low-bandwidth connections. The SPS takes directives from the VPA modules to instantiate specific detections of interest.

The low-latency user location information generated by the SPS and VPA modules on AR-Mentor interactions are forwarded to the Rendering subsystem. The Rendering modules generate animations that exactly match the users perspective view as overlays in the OST HMD. The VPA generated textual phrases are also converted to speech for auditory feedback to the user.

3 TECHNICAL APPROACH

Pose estimation module is used to estimate the 6-DOF pose (3D location and 3D orientation) of the user with respect to the equipment that is being maintained. It consists of the following three tightly coupled blocks. (1) Landmark Matching Block: We developed a database of visual features of the training vehicle. This module establishes the position of the user with respect to the vehicle; (2) Visual Odometry and Navigation Module: This block tracks visual features over time to know how the user's head is moving in real-time. These features with IMU sensor measurements and Landmark data are integrated in the filter to get precise 3D head position and orientation of the user. (3) Low-Latency Prediction Module: This uses information from the block above along with the IMU to predict where the user would look at the exact time when the information is displayed.

3.1 High Precision Localization

We employ an IMU-centric error-state Extended Kalman Filter (EKF) approach [12] to fuse IMU measurements with external sensor measurements that can be local (relative), such as those provided by visual odometry, or global, such as those provided by visual landmark matching. The filter replaces the system dynamics with a mo-

tion model derived from the IMU mechanization model which integrates the incoming gyro and accelerometer readings to propagate the system state from a previous frame to the next. Process model follows from the IMU error propagation equations, which evolve smoothly and therefore are more amenable to linearization. This allows for better handling of the uncertainty propagation through the whole system. The measurements to the filter consist of the differences between the inertial navigation solution as obtained by solving the IMU mechanization equations and the external source data. At each update, the EKF estimated errors are fed back to the mechanization module to not only compensate for the drift that would otherwise occur in unaided IMU but also to correct the initial conditions for data integration in the mechanization module. Figure 4 shows the core blocks that make up the localization system.

Figure 4: Localization flowchart.

Our system uses vision algorithms for both relative pose computation and absolute pose computation. These are both done as inputs in terms of feature based image correspondences to the Kalman filter. In our EKF framework, we use both relative measurements in a local 3D coordinate system via visual feature tracks and absolute measurements via 3D-2D landmark tie-points as inputs. We compute a 6 DOF pose (both 3D rotation and 3D translation). The visual feature track measurements are applied in a strictly relative sense and constrain the camera 6-DOF poses between frames. Each feature track is used separately to obtain its 3D position in a local coordinate system and a measurement model whose residual is based on its re-projection error in the current frame is used to establish 3D-2D relative constraints on the pose estimate. The 3D location for each tracked point is estimated using all frames in which it was previously observed and tracked. Simultaneously, 3D-2D measurements arising from landmark matching are fed to the filter directly and used in an absolute sense for global geo-spatial constraints. Within this framework, the navigation filter can handle both local and global constraints from vision in a tightly coupled manner, and our localization system is able to accurately track the user's head location within 5-centimeter errors over hours of usage [21].

3.2 Landmark Database Building

The landmark matching module correlates what the user is seeing with a pre-created visual landmark database to locate the user with respect to the target training objects. The landmark matching module is divided into two sub-modules: landmark database creation and online matching to the pre-created landmark database. During landmark database creation, a set of video sequences are collected using stereo sensors. For example, for the training vehicle case, from the collected video sequences, an individual landmark database was created for different key locales on the vehicle. Each individual landmark is characterized by a unique Locale ID and Object State ID. Locales include the Turret, Cargo Bay, etc. The state ID's include detections such as hatch open/close, shield on/removed, etc. Then we collect and categorize them into locales and states and create a

single visual database for the vehicle. This database is used for all subsequent training events.

During a live maintenance procedure we extract landmarks from the live video and match them to the pre-built landmark database. If a match is found, the Locale ID is returned. Given the returned Locale ID information, the pre-built landmark database can be further constrained or narrowed for the next input query images to obtain both faster and more accurate positioning of the user and states.

Figure 5 shows a set of randomly selected sample images for building the training vehicle landmark database. Some details of building such a landmark database is given in the following sections.

Figure 5: A set of randomly selected landmark images of the training vehicle during the landmark database creation.

3.2.1 Multi-stage Bundle Adjustment Alignment

The equipment usually consists of several movable or moving parts. For example, for the training vehicle, as shown in Figure 6, the launcher can be moved up and down, and the cover of the switch needs to be removed in order to see the inside during the maintenance procedure. Therefore, the issue is how to build a consistent landmark database for the equipment with various removable and articulated parts.

Figure 6: Sample images of the training vehicle: (Left) the launcher down; (Middle) the launcher up; (Right) the launcher up and the switch cover off.

For the training vehicle, since it is a large object (Length, 6.45 m; Width, 3.20 m; Height, 2.97 m), during the data collection stage, we need to collect all the possible scenarios with moving parts, such as launcher is up, launcher is down, the switch cover is off, etc. Once the video data is collected for all cases, we need to align them together using the bundle adjustment algorithm [18]. After bundle adjustment, the whole vehicle is reconstructed from selected key-frames and all articulated vehicle parts are aligned automatically into a globally consistent coordinates system.

During online landmark matching, since the underlying image matching algorithm [21] utilizes Histograms of Oriented Gradients (HOG) features that are invariant to lighting changes in images, vehicles of the same type are matched under various indoor/outdoor environments. In fact, only landmarks of the vehicle are actually useful for matching, and landmarks of the surroundings serve as outliers and degenerate matching performance. Therefore, landmarks of surroundings have to be removed from the landmark database.

3.2.2 Automatic Background Removal

From the reconstructed 3D sparse point cloud of the whole vehicle with its surroundings, we are able to manually segment out the vehicle as shown in Figure 7.

With a manually segmented 3D vehicle, we can simply define the 3D regions of the vehicle in the point cloud, and use the defined

Figure 7: The segmented 3D reconstructed sparse point cloud of the vehicle: left: front-side view; right: overhead view.

3D vehicle regions to remove all 2D background landmarks in the scene automatically for each 2D key-frame image or selected landmark shot. Finally, the size of the built vehicle landmark database is 419 MB, and it contains 6564 landmark shots.

3.3 Low Latency Prediction Module

For OST AR, accuracy of the pose estimates alone is not sufficient for delivering an acceptable user experience to the person who is wearing AR-Mentor. For example, besides rendering a virtual marker at the correct location, the rendered marker also needs to appear with very little delay on the display. This is due to the fact that, in the OST framework, the user sees the real work as it is (not an image of it) and hence the equivalent frame-rate is essentially very high and there is no-delay in visual perception of the real world. Therefore, the associated rendered markers have to satisfy this highly demanding requirement in order for them to appear jitter-free when they are displayed. Otherwise as the users head moves, the markers will appear to bounce around in the display since they will be lagging in time.

Video frames in general arrive (15 Hz in our case) at a much slower rate than the IMU samples (100 Hz in our case.) The pose estimates that incorporate each video frame information is generally available after a 40-50 msec processing delay. The pose requests from the renderer arrive asynchronously at the highest rate the renderer can accommodate. After the renderer receives a pose it is displayed on the see through display after a certain amount of delay which is affected by both the display hardware latency and lag caused by the inefficiencies in the rendering pipeline and video graphic card. In order to compensate for all the latencies in the system, a forward prediction mechanism is utilized to estimate the camera pose corresponding to a certain timestamp into the future given all the information that is available up until the render request. For this purpose, forward prediction performs a second-order extrapolation of the orientation using a window of past camera poses with the Kalman Filter [12].

4 VIRTUAL PERSONAL ASSISTANT (VPA)

Successful task execution requires a rich dialogue interaction between the user and the system. VPA is a conversational computerized personal assistant that guides the user in performing complex tasks [2, 6, 14, 3]. VPA has a deep understanding of the domain and is able to guide the user through every step by constantly monitoring and understanding the users context [4]. VPA is designed to robustly exploit the context of the interaction in formulating its responses. This gives it the capability (among others) to hold a true back-and-forth conversation with the user, not merely respond to one-shot tasks or queries.

4.1 VPA Components

As shown in Figure 8, the user interacts with the VPA system through multimodal inputs, such as natural language speech/text; gesture; user interface; vision; or other sensory input. In the AR-Mentor system, speech is converted into text before it is passed to VPA. In addition, VPA also gets the location of the user and what the user is looking at. All the multimodal input signals are normalized and

converted into user's goals. The understanding component combines rule-based techniques and statistical models to come up with the candidate intents that represent VPA's understanding of what the user wants to do at that moment. Intents in VPA are captured by semantic frames that represent actions, with slots representing the semantic entities associated with those actions. The rule-based parser is based on the Phoenix semantic parser [16]. The rule-based parser captures the correlation between syntactic patterns and their correlated intent semantics by means of manually constructed rules. The statistical models consist of MaxEnt (maximum entropy) models for intent classification and argument classification [11]. Given an utterance, the intent classifier identifies the most likely intent (frame) and the associated intent-specific argument classifiers predict the most likely arguments (slot values). The statistical models are learned from labeled training data obtained from subject matter experts. VPA relies on domain-specific heuristic techniques to determine the final intent, based on outputs from the rule-based parser and the statistical models.

Figure 8: VPA system overview.

A context-sensitive interpreter is used to evaluate all the candidate intents in terms of the current conversational and operational context and to come up with a final user's intent that is passed to the reasoning component. The understanding component can also extract useful information about the user's state and characteristics, which could enable adapting and customizing the user interaction. End users can change intents during their interaction with VPA, and return to previous intent completion in a natural way.

The reasoning component executes the best course of action by using the business rules and the domain-specific knowledge through deep integration with the backend business logic and data, allowing users to perform complex functions (multi-part transactions, setting triggers, etc.). The VPA platform supports this integration through a documented implementation methodology and a graphical workflow tool for rapidly creating and modifying complex workflows. The workflow includes the knowledge coded in the technical manual such as instruction for executing the steps, safety information, control flow, parts and tools check, preconditions, knowledge gathered from an experienced instructor such as practical tips, answers to typical user questions, and control scripts for managing the flow based on user's context. The reasoning component leverages knowledge about the user, models of domain products, and descriptions of domain processes to determine an appropriate action by the system, either in response to user queries or by proactively initiating a dialog. The reasoning component creates an output intent based on its decision what to do next and passes it on to the output module.

The output intent determines the best modality and format to communicate to the user at the moment. The output intent is converted into multimodal information for playback to the user in natural language speech/text and/or via other sensory outputs, such as user interface, AR, etc. In the AR-Mentor system, the output module can come up with a sequence of video segments to play, the images to overlay and the audio to play related to that action. The output module component uses domain-specific templates to build answers from the output intent. The output module supports separate templates for

text, speech, and UI modes. The dialog cycle continues until the goal is accomplished or the user decides not to proceed.

4.1.1 Spoken Natural Language Interaction

The Automatic Speech Recognition (ASR) module converts speech to text [19, 10] that is being passed to the VPA. AR-Mentor uses DynaSpeak [1] as its ASR and is a commercial software developed by SRI International. DynaSpeak is a high accuracy speaker independent speech recognition engine that scales from embedded to large-scale system use in industrial, consumer, and military products and systems. DynaSpeak supports both finite state grammars - used in more traditional command and control style applications - and statistical language models - used in more advanced natural language style dialog applications. DynaSpeak uses standard acoustic models that have cross-word triphones modeled as hidden Markov models with output distributions modeled by Gaussian mixture distributions trained discriminatively using the minimum phone error criterion. The front-end signal processing uses Mel-frequency cepstral coefficients features, which are transformed using several linear operations to adapt to speakers and environments. To speed up ASR decoding, DynaSpeak uses standard Gaussian selection shortlists. The ASR decoding graph is created using a unigram language model (LM) using highly optimized weighted finite state transducer composition and expanded using a modified on-the-fly LM rescoring approach with a 4-gram, Kneser-Ney smoothed LM. Detailed information about the ASR system can be found in [20].

For better speech-recognition performance, the ASR component must be updated with language and acoustic models for the specific domain. This refinement can be done by collecting existing audio data, gathering audio data during customer interaction with the system, and gathering textual data about the domain. AR-Mentor mainly uses textual data to build the domain-specific models. The AR-Mentor environment requires the system to be operated hands-free and allow the user to communicate with the helper at any time. The ASR was updated to handle the continuous speech. It would then leverage the target phrase and a pause in the speech to determine the request from the user.

```

...
AR-Mentor: What's wrong with your training vehicle?
User: The up indicator light won't come on.
AR-Mentor: OK. Sounds like you need to adjust the upper switch. Let's start by checking that you have the required tools
for the work. Do you have your multi-capable maintainers' tool kit?
User: Yes, I have that.
AR-Mentor: Do you have a socket adapter, one-half inch to three-eighths inch?
User: What is that used for?
AR-Mentor: The socket adapter is needed to attach the three-eighths inch hex key adapter to the one-half inch drive torque
wrench to properly tighten the hex screws in the shield assembly. Say 'yes' when you have one.
User: Yes.
...
AR-Mentor: The next few steps require you to be in the rear hatch. Move to the rear hatch and let me know when you're there.
<User enters the rear hatch area>
AR-Mentor: I can see that you're in the rear hatch now, so let's continue with the steps. Before we do the next step, please pay
attention to this important Safety Note: An unsupported metal lid can fall unexpectedly and kill or seriously injure
you. Do not work under an unsupported metal lid. Understood?
User: Understood.
AR-Mentor: All right, with that said: Step 10 is to remove the shield from the housing. You do that by removing the four screws,
and then the shield, from the housing. A tip: This shield is deceptively heavy, and it'll smash your fingers if you're not
careful. Be sure to have your helper assist you, and lift the shield from the sides, not the bottom. OK, remove the shield
and let me know when you're done.
User: Why are we doing this?
AR-Mentor: The upper switch that we're adjusting is in a gearbox that's behind this protective shield. We need to remove the
shield to access the gearbox.
<User removes the shield>
User: done.
AR-Mentor: The next step also comes with an important Safety Warning: The lid, in the up position, can fall rapidly and
injure personnel! Stay clear of the metal lid path when the metal lid is up. Understood?
User: Yes, understood.
AR-Mentor: Keep that in mind when doing Step 11: Remove plug from jack. Let me know when that's done.
User: Where is that again?
AR-Mentor: The plug is located on the rear of the gearbox, connected to the upper switch cover. I'm pointing you to it.
...

```

Figure 9: Example of Natural Language dialogue of the AR-Mentor had with the user.

Figure 9 shows sample snippets of dialog between the user and AR-Mentor during live training. The first few exchanges show examples from the preliminary phases of the work package training: those verifying that the user has all the required tools and parts, and that the equipment is in the appropriate starting condition. Later exchanges show examples from steps in the actual maintenance. These include automatically verifying that the user is in the right location to perform the upcoming steps, providing the user with important safety warnings and tips, and describing the steps themselves. The system gives informative answers to user questions about precise locations of important parts, the purpose of steps and tools, and other topics. In addition, the system interprets those questions in the context of the conversation; for example, when the user asks "What is that used for?", the system infers from context what "that" is. Each AR-Mentor utterance, including answers to user questions, is accompanied by an animation or image that illustrates the concept being taught.

5 SYSTEM SET-UP

The AR user-worn hardware is used by the system to both sense the environment using audio, visual and inertial measurements and to output guidance to the user through natural language spoken dialogue and visual cues augmented on the user's HMD. A simple, lightweight sensor/processor package is essential for the user to observe and manipulate the objects in front of him freely and naturally. For the purpose of computational complexity and scalability, we ran the AR-Mentor system in client-server mode as shown in Figure 10.

Figure 10: The client-server model of the AR-Mentor.

As shown in Figure 10, the non-time-critical processing components such as VPA and ASR is running on a standalone remote laptop server and all the time-critical processing tasks including user head localization and HMD rendering are running on a compact Apple Mac-Mini computer (2.3GHz Quad-Core i-7 processor) that the user is wearing. The two systems communicate wirelessly. In addition, an Li-ion battery is carried with the user to power the Apple Mac-mini machine and sensor-rig and HMD. It weighs around 3 lbs and can last for three hours. All equipment is carefully configured to fit a vest with MOLE straps as shown in Figure 10.

5.1 The 33-Step Maintenance Task

We have built a pilot application using AR-Mentor to help student mechanics learn and perform a complex and advanced maintenance procedure for the training vehicle. It consists of 33 steps covering 11 written pages in the technical manual, and it usually takes approximately 40 minutes for an experienced training vehicle mechanic to perform. The mechanics training school (GA, US) currently devotes up to 9 hours to groups of about 30 mechanics to practice this procedure hands-on. It requires substantial monitoring by the instructor to perform the task properly. It is a very high burden on the instructor. AR-Mentor is designed to help relieve this burden.

Figure 11 shows a few example images with virtual insertions for step 6 and 10, where step 10 instructs the student mechanics how to

Figure 11: A set of selected example images of step 6 (upper row) and step 10 (lower row) with virtual insertions (tools, parts and texts).

use a ratchet wrench to remove the shield from housing by removing four screws in necessary order.

6 CONCLUSION AND FUTURE WORK

In this paper, we described a highly flexible, mobile, automated mentoring and training system, named AR-Mentor, tailored to the needs of individual users. The system acts as a personal mentor to a user, providing human-like understanding and guidance. It will go where the user goes, from classroom to battlefield. It interacts with the user in natural spoken language and through visual (AR) indicators in the scene. It provides training in both the operation and maintenance of equipment. The current AR-Mentor prototype is the first mentoring system that uniquely combines AR, visual and natural language understanding and reasoning and virtual personal assistance technologies together. Preliminary live training exercises with real novice users using AR-Mentor demonstrated promising effectiveness of helping users to get the task done successfully.

In the future, we will improve AR-Mentor as follows. First, we will focus on the reduction of both size and weight of the sensor-rig on the user's head. We will replace the stereo-cameras with a single monocular camera and update the localization algorithm with the monocular camera accordingly. Second, we will reduce the weight of both computer and batteries on the user's body by offloading the time-sensitive processing on the current Mac-Mini to a smart-phone package with mobile processors, which weighs much less and consumes much less power. Finally, on the VPA side, we will focus on simplifying the authoring process for procedural tasks and injecting more robust diagnostic reasoning into VPA for smarter diagnostic training.

ACKNOWLEDGEMENTS

This material is based upon work supported by U.S. Army Project: Augmented Reality based Training (AR-Mentor) under Contract W91WAW-12-C-0063. The views, opinions, or findings contained in this report are those of the authors and should not be construed as an official Department of the U.S. Army position, policy, or decision unless so designated by other official documentation.

REFERENCES

- [1] <http://www.speechatsri.com/products/dynaspeak.shtml>.
- [2] P. Berry, M. Gervasio, B. Peintner, and N. Yorke-Smith. PTIME: Personalized assistance for calendaring. *ACM Transactions on Intelligent Systems and Technology*, 2(4), 2011.
- [3] P. Berry, K. Myers, T. Uribe, and N. Yorke-Smith. Task management under change and uncertainty: Constraint solving experience with the CALO project. In *Workshop on Constraint Solving under Change and Uncertainty*, 2005.
- [4] H. Bui, F. Cesari, D. Elenius, D. Morley, S. Natarajan, S. Saadati, E. Yeh, and N. Yorke-Smith. A context-aware personal desktop as-

sistant. In *International Joint Conference on Autonomous Agents and Multiagent Systems: Demo Papers*, 2008.

- [5] D. Curtis, D. Mizell, P. Gruenbaum, and A. Janin. Several devils in the details: making an ar application work in the airplane factory. In *International Workshop on Augmented Reality (IWAR)*, 1998.
- [6] W. Haines, M. Gervasio, A. Spaulding, and B. Peintner. Recommendations for end-user development. In *ACM Workshop on User-Centric Evaluation of Recommender Systems and their Interfaces*, 2010.
- [7] S. J. Henderson and S. Feiner. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *International Symposium on Mixed and Augmented Reality (ISMAR09)*, pages 135–144, 2009.
- [8] Y. Kim and I. Moon. E-training content delivery networking system for augmented reality car maintenance training application. *International Journal of Multimedia and Ubiquitous Engineering*, 8(2), 2013.
- [9] K. Lee. Augmented reality in education and training. *TechTrends*, 56:13–21, 2012.
- [10] V. Mitra, M. McLaren, H. Franco, M. Graciarena, and N. Scheffer. Modulation features for noise robust speaker identification. In *Proc. of Interspeech*, 2013.
- [11] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, ISMAR '06, 1999.
- [12] T. Oskiper, H. Chiu, Z. Zhu, S. Samarasekera, and R. Kumar. Stable vision-aided navigation for large-area augmented reality. In *IEEE Virtual Reality (VR)*, 2011.
- [13] T. Oskiper, M. Sizintsev, V. Branzoi, S. Samarasekera, and R. Kumar. Augmented reality binoculars. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2013.
- [14] B. Peintner, J. Dinger, A. Rodríguez, and K. Myers. Task assistant: Personalized task management for military environments. In *Conference on Innovative Applications of Artificial Intelligence (IAAI-09)*, 2009.
- [15] J. Platonov, H. Heibel, P. Meier, and B. Grollmann. A mobile markerless ar system for maintenance and repair. In *International Symposium on Mixed and Augmented Reality*, ISMAR '06, 2006.
- [16] W. Ward. Extracting information from spontaneous speech. In *International Conference on Spoken Language Processing*, 1994.
- [17] S. Weibel, U. Bockholt, T. Engelke, N. Gavish, M. Olbrich, and C. Preusche. Augmented reality training for assembly and maintenance skills. *Robotics and Autonomous Systems*, 61(4):398–403, 2013.
- [18] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *CVPR*, June 2011.
- [19] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang. Automatic phonetic segmentation using boundary models. In *Proc. of Interspeech*, 2013.
- [20] J. Zheng, A. Mandal, X. Lei, M. Frandsen, N. F. Ayan, D. Vergyri, W. Wang, M. Akba-cak, and K. Precod. Implementing sri's pashto speech-to-speech translation system on a smartphone. In *IEEE Workshop on Spoken Language Technology (SLT)*, 2010.
- [21] Z. Zhu, H. Chiu, S. Ali, R. Hadsell, T. Oskiper, S. Samarasekera, and R. Kumar. High-precision localization using visual landmarks fused with range data. In *IEEE Conference on CVPR*, 2011.