Augmented Reality Virtual Personal Assistant for Training, Maintenance, and Repair

Rakesh (Teddy) Kumar, Supun Samarasekera, Girish Acharya, Louise Yarnall, Zhiwei Zhu, Michael Wolverton, Vlad Branzoi, Glenn Murray, Nicholas Vitovitch, Ryan Villamil, Jim Carpenter
SRI International
Princeton, NJ
rakesh.kumar@sri.com, supun.samarasekera@sri.com

ABSTRACT

The military trains a large pool of personnel skilled in maintaining and repairing a variety of complex equipment. The U.S. Army itself requires personnel for more than 130 different Military Operational Skills. Often these trained personnel are not available for repair and maintenance of critical low density equipment in deployed locations. Augmented Reality and Virtual Personal Assistance are technologies that can supplement live training to address the challenge of affordably training personnel.

In this paper, we present the system design, hardware, algorithms and initial field results for a prototype training system AR-Mentor. The system is designed to act as a personal mentor to a user, providing human-like understanding and guidance. It provides a Heads-up and Hands-free experience. The user can train anywhere and also use the system for providing guidance during actual maintenance of the equipment.

The experimental system consists of a compact computer, head worn cameras, microphone, ear-buds and eyewear. Virtual Personal Assistant technology is used to provide a real-time dialog and reasoning system that supports human-like interaction using spoken natural language. The reasoning system aims to recognize the user's intent and provides feedback to the user. The feedback and interaction occurs both verbally and by engaging the Augmented Reality system to display icons and instructions visually on the user's eye-glasses. The inserted visual objects appear as part of the live scene and are precisely aligned to the equipment.

A formative evaluation indicated that the AR-Mentor system permitted individual learners to focus on their learning needs and reduced the perceived mental demand of learning the procedure. Checks into understanding showed no difference between learning with the AR-Mentor system, as compared to learning from an instructor, or a technical manual. The evaluation also indicated the need for alternative ways to design the AR-Mentor representations around complex procedural steps.

ABOUT THE AUTHORS

Dr. Rakesh "Teddy" Kumar is the Director of the Center for Vision Technology at SRI International, Princeton, NJ. Prior to joining SRI, he was employed at IBM. He received his Ph.D. in Computer Science from the University of Massachusetts at Amherst in 1992. His technical interests are in the areas of computer vision, computer graphics, image processing and multimedia. Rakesh Kumar received the Sarnoff Presidents Award in 2009 and Sarnoff Technical Achievement awards in 1994 and 1996 for his work in registration of multi-sensor, multi-dimensional medical images and alignment of video to three-dimensional scene models, respectively. He was an Associate Editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence from 1999 to 2003. He has served in different capacities on a number of computer vision conferences and National Science Foundation review panels. He has co-authored more than 50 research publications and has received over 35 patents.

Mr. Supun Samarasekera is the Technical Director of the Vision and Robotics lab, Center for Vision Technologies at SRI International, Princeton, NJ. He received his M.S. degree from University of Pennsylvania. Prior to joining SRI, he was employed at Siemens Corporation. Supun Samarasekera has 15+ years' experience in building integrated multi-sensor systems for training, security & other applications. He has led programs for robotics, 3D

modeling, tr`aining, visualization, aerial video surveillance, multi-sensor tracking and medical image processing applications. He has received number technical achievement awards for his technical work at SRI.

Mr. Girish Acharya is an Engineering Director at SRI International, Menlo Park, CA, responsible for integrating technologies and the product transition to the customer. He has a Master of Applied Science from University of Toronto and Master of Business Administration from Haas School of Business at U.C. Berkeley. Mr. Acharya has more than twenty years of managing technical projects and is currently managing Virtual Personal Assistant project. He was a member of the senior management team for SRI's CALO project in the DARPA PAL program as well as in DARPA MALTA program to transition PAL technologies to the U.S. military.

Dr. Louise Yarnall is a Senior Research Social Scientist in assessment development and cognitive task analysis. She has a Ph.D. in Education with specialization in educational psychology and technology in learning, UCLA. Dr. Yarnall has more than 12 years' experience designing technology-based educational materials and assessments. She has served as principal investigator and evaluator on grants for the National Science Foundation and the U.S. Department of Education's Institute for Education Sciences. She is an expert in employing evidence-centered design (ECD) to design and develop assessments.

Dr. Zhiwei Zhu is a Senior Computer Scientist at SRI International, Princeton, NJ. He has a Ph.D. in Electrical Engineering from Rensselaer Polytechnic Institute, Troy, NY. His main research focus is in the area of Computer Vision and Human Computer Interaction. He has published over 40 journal and conference papers, and has received one Best Transaction Paper Award from IEEE Transactions on Vehicular Technology in 2004 for his driver fatigue monitoring work and another Best Paper Award at IEEE Virtual Reality Conference in 2011 for his co-authored work in the high-precision localization and tracking for the large-scale infrastructure-free augmented reality applications.

Dr. Michael J. Wolverton is a Senior Computer Scientist in SRI's Artificial Intelligence Center, Menlo Park, CA, where for the past 15 years he has led or played key technical roles in research projects in automated personal assistance, planning, explanation, link discovery, intelligent information management, and other areas. Recently he has led the Reasoning task within SRI's Virtual Personal Assistant effort, where he has developed approaches to dialog management and acquisition of dialog knowledge. He holds a bachelor s degree in Mathematical Sciences and Computer Science from Rice University, and a Ph.D. in Computer Science from Stanford University.

Mr. Vlad Branzoi is a Senior Computer Scientist at SRI International, Princeton, NJ. He received his M.S. in Computer Science from Columbia University under Prof. Shree Nayar. Vlad Branzoi has over 10 years' experience in building novel sensors, integrated multi-sensor systems for training, robotics and mobile applications.

Mr. Ryan Villamil is a Senior Computer Scientist at SRI International, Princeton, NJ. He received his MS in Computer Science from Columbia University. Ryan has specific technical expertise in computer graphics, computer graphics, simulation and training system. He has developed high performance rendering, visualization and gaming systems for a number of mixed and augmented reality applications.

Mr. Nicholas Vitovitch is an Associate Computer Scientist with the Modeling and Robotics group at SRI International, Princeton, NJ. He received his B.S. degree from The College of New Jersey in both computer science and mathematics where his research interests included computer vision and cloud computing. His academic pursuits were recognized by his induction into the Upsilon Pi Epsilon and Phi Beta Kappa academic honor societies. Since joining SRI he has contributed to the development of augmented reality training and simulation systems, aerial LIDAR classification and modeling tools, and mobile surveillance and multi-sensor tracking applications. His current interests include deep learning and wearable technologies.

Mr. Glenn Murray is a Computer Scientist at SRI International, Princeton, NJ. Glenn has over 30 year's software development experience, 20 years as an independent software development consultant. Glenn has technical expertise in many disciplines including system engineering, large-scale multi-threaded/multi-processor systems, communications and graphical user interface.

Augmented Reality Virtual Personal Assistant for Training, Maintenance, and Repair

Rakesh (Teddy) Kumar, Supun Samarasekera, Girish Acharya, Louise Yarnall, Zhiwei Zhu,
Michael Wolverton, Vlad Branzoi, Glenn Murray, Nicholas Vitovitch, Ryan Villamil, Jim Carpenter
SRI International
Princeton, NJ
rakesh.kumar@sri.com, supun.samarasekera@sri.com

INTRODUCTION

The military trains a large pool of personnel skilled in maintaining and repairing a variety of complex equipment. The U.S. Army itself requires personnel for more than 130 different Military Operational Skills. These trained personnel are critical for repair and maintenance of critical low density equipment in deployed locations. It usually takes months, even years, of hands-on training to become a "skilled" professional mechanic. Additionally, there is simply not enough time in schoolhouses to teach every possible maintenance procedure even when the scope is limited to a single line of complex machines, much less an ever evolving one. Mechanics must continuously train on and learn new technologies and procedures. Given these circumstances, developing technologies to improve the efficiency and effectiveness of instruction and to support in-the-field, on-the-job assistance for new tasks (never-seen-before) becomes a very important exercise.

Recent research suggests that Augmented Reality (AR) and Virtual Personal Assistant (VPA) technologies offer the potential to achieve improved learning and on-site human-instructor-like guidance. When performing complex tasks, having in-context visual and verbal guidance through the task enables faster and more efficient repair. This can be used both for training and in-the-field operations. AR provides an overlay of real-time visual information on a user's view of the physical world, to guide him in performing tasks; while VPA supports human-like interaction using spoken natural language and recognizes the user's goals and provides feedback to the user in real- time. Together, these technologies 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. In addition, these technologies have the potential to: improve learning and long-term memory through parallel processing of multiple representations of information (e.g., video, imagery, animations and audio) [Mayer 2002]; reduce the demands on learners' short-term memory during learning by

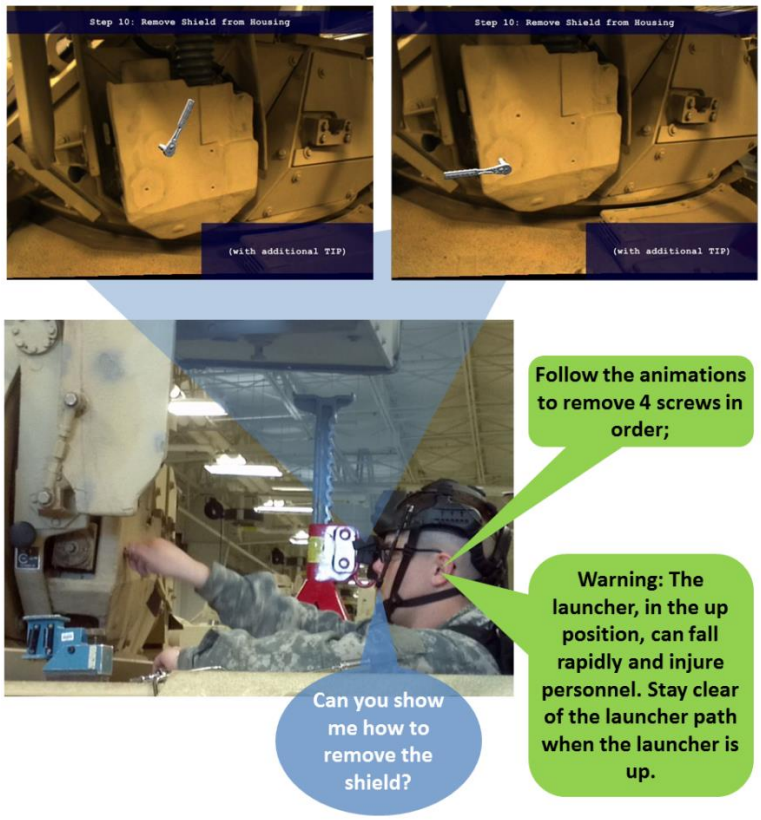

Figure 1: Concept of AR-Mentor System. The user is able to speak to the AR-Mentor system (via Micro-phone); The AR-Mentor system understands the user and gives back the instructions both verbally (via Speakers) and visually (via Optical-See-Through Glasses).

presenting dynamic representations of multi-step procedures [Chandler 2009, Hegarty 2003]; focus learner attention efficiently on the central elements of a complex procedure [Yantis 1990]. Additionally, such technologies have the potential to increase the time in the schoolhouse for higher-order learning. For example, they may reduce the burden on instructors for low-level, redundant teaching activities such as correcting errors in technical manuals and re-explaining basic procedures, thus freeing up the opportunity for instruction around higher-order problem solving. They may also reduce the burden on peer "helpers" who read technical manuals aloud, thus increasing the opportunity for learners to engage in more complex activities, such as collaborative problem solving.

Based on these principles, we have developed a prototype Augmented Reality-based Mentoring system (AR-Mentor). We present the technical details and experimental results of the AR-Mentor system in this paper. As shown in **Figure 1** above, it is a head-worn optical-see-through display device through which technicians can receive just-in-time audio-visual guidance while learning to conduct adjustments/repairs on the equipment. The AR-Mentor also provides guidance using on-demand voice instruction and five types of visual overlays to the work environment: 3D graphic animations to describe tools and components and how to manipulate them, animated arrows that direct the learner' gaze direction, live-action videos of maintainers conducting adjustment procedures, text-annotated graphic images of complex tools, and diagrammatic images of complex equipment and the vehicle location map. The learner can interact with the system, ask questions and may direct the flow of the 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 on the equipment to conduct an adjustment and provides feedback on correct placement.

PREVIOUS WORK

Currently, the most commonly used mentoring systems in training are either pure simulation or interactive voice response (IVR) systems. Pure simulation systems, such as desktop simulations or virtual reality systems, while providing an enhanced version of an instruction manual are by definition hands-off. Many of these systems lack the physical feedback provided by hands-on experience with complex pieces of equipment that enable comprehensive training. In IVR systems, the system directs the dialogue, and the user is constrained to a maze of questions and limited answers. AR, on the other hand, is able to blend 2D or 3D virtual overlays onto the direct view of the real world of the students, which allows the students to practice live training exercises repeatedly in a realistic AR-enhanced environment. As a result, various AR-based training systems [Kumar_2012, Henderson_ETB2009, Webel_ART2013, LeeAR2012, KimETC2013, Oskiper 2013] have been proposed in the past few years. However, as pointed out by Lee [LeeAR2012], 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. [Kim ETC 2013] 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. [Henderson_ETB2009] presented a prototype AR application to support military mechanics conducting routine maintenance tasks inside an armored vehicle turret. Custom-built stereo Video See-Through (VST) Head-Mounted Display (HMD) is used instead of Optical-See-Through (OST) HMD, and a wrist-worn controller is used to control the animations or cue the next task, which is not hands-free. In addition, they installed 10 tracking cameras around the turret and installed three infrared (IR) light emitting diodes (LEDS) on the HMD to track the user' head movement. Experiments showed that AR helped mechanics to locate tasks more quickly than when using both baseline conditions without AR. Webel et al. [Webel_ART2013] 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. Their experiments concluded that AR has the potential to be a useful technology for maintenance and assembly training and AR-based training does not require an on-site trainer.

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 will have 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 even in theater.

The main goal of the paper is to present the system design, algorithms and formative evaluation of an OST-AR and VPA-based mentoring system. We have successfully developed an experimental AR-Mentor system that can be used by the student mechanics to learn and perform a 33-step vehicle maintenance task without any technical manuals or instructors. It can be configured easily to assist any other maintenance and repair tasks of vehicles, or any other complex machinery. It consists of a user worn OST display eye-wear and head-phones that can (a) Talk to the user to give directions and guidance, (b) Display textual information of tasks and (c) Overlay symbolic icons and directions that precisely match the vehicles parts that are being observed. To do so it is important for the system to understand the task context. Context is obtained through (a) having a microphone system that can listen to the user's speech and (b) a video-based sensor package that can accurately locate the user with respect to the vehicle and interpret his actions. The system operates in a hands-free and heads-up mode with natural language spoken interactions ensuring uninterrupted task attendance while learning. Finally, formative evaluation observations were conducted to gather early-stage evidence of learner satisfaction using AR-Mentor and behavioral indicators for improving the design of the AR-Mentor system.

TECHNICAL APPROACH

The AR-Mentor system aims to mimic the hands-on interactive experience that a one-on-one individual trainer or mentor would provide a user to train them to accomplish complex tasks. The goal is to provide guidance in both the operation and maintenance of the user's equipment. The AR-Mentor system provides human-like understanding and guidance accompanying the user from classroom to battlefield, interacting with the user in natural spoken language and through visual (AR) indicators in the scene.

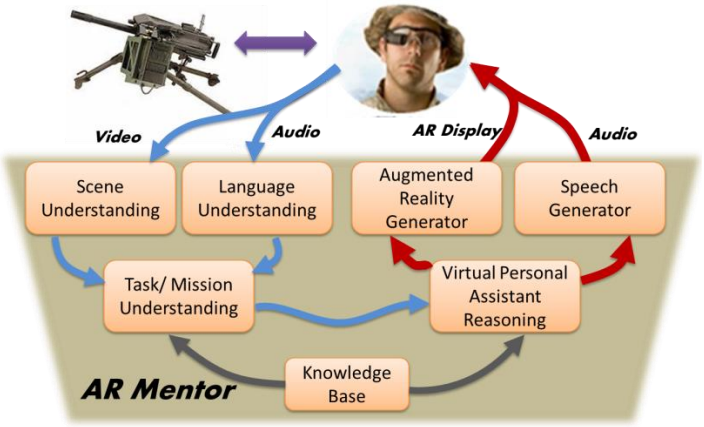

Figure 2: AR-Mentor System: User Worn AR, VPA and Mentoring System. Audio/ Visual cues are used by the system to understand the dynamic scene and the user's intent with respect to the task being performed. Guidance is provided by the VPA Reasoning module in terms of natural language spoken dialogue and visual cues augmenting the user's HMD live view.

Figure 2 outlines a high-level system work-flow diagram of the AR-Mentor system a user wears a head-mounted sensor/display package to use AR-Mentor. The system uses visual cues (from video) to immediately situate the user with respect to the world, including any equipment the user is being trained on. The exact relative position and head orientation of the user is tracked continuously by the system. The ***Sensor Processing*** module uses these cues and observed scene characteristics to understand user action and intents. The scene understanding is augmented further by the ***Language Understanding*** module. This module uses speech recognition as a front-end to interpret the language-based inputs provided by the user. The ***Task Mission Understanding*** module combines inputs from Scene Understanding and Language Understanding and uses the ***Knowledge Base*** to understand the user's current state and intent. A core component of the AR-Mentor system is the ***VPA Reasoning*** module**.** This module interprets the intent cues generated from the audio-visual understanding modules in the context of the task ontologies and workflow models. The system is able to then reason about the next step in an interactive dialog that the system needs to conduct with the user to achieve the goal. We use hierarchical action models to define tasking cues relative to the workflow ontologies that are defined. The output of the reasoning module is fed to ***Augmented Reality Generator/3D Animation System*** to create display content that takes the world model and user's perspective into account. Similarly output of the reasoning module is used by the ***Speech Generator*** module to create context dependent verbal cues. The system talks to the user while displaying the corresponding animations that are precisely aligned with the world the user is seeing.

Scene and Language Understanding

The function of the understanding block is to take low-level sensor data (audio, visual and inertial) and determine intent of a collaborative user in the context of well determined workflow for performing a complex task. As the user

performs the task and progresses through the workflow, user intents are automatically generated by the understanding block and are fed to a reasoning engine that determines the audio-visual guidance to be provided at the next instant. Starting with audio-visual input and making a determination of intent is challenging. As such the understanding modules form a significant core collection of advanced capabilities we have developed and plan to extend.

Scene Understanding: Localization

AR training systems using HMDs require high precision knowledge of the 3D location and 3D orientation of the user's head. This is required by the system to know where to insert the synthetic actors and objects in the HMD. Inserted objects must appear stable and not jitter or drift with respect to the real world. Moreover latency of less than 5 milliseconds for pose estimation is required for lag-free see-through HMD operation. We achieve this performance using a multi-sensor navigation system mounted on the HMD. We exploit a head mounted sensor package that consists of a camera, inertial measurement unit (IMU) and optionally GPS unit.

For mentoring applications (Training or Assistive) the objects of interest (or the locale) are well defined. In such case the visual features of the object (or locale) can be extracted in advance for providing positioning with respect to the object in real-time. The landmark matching/object recognition module allows us to pre-build a landmark/object database of the objects/locales and use it to define users' movements relative these objects/locales. Using our helmet, we collect imagery and 3D data to build 3D models and landmark databases of the objects of interest. We can also take advantage of any pre-built 3D models that might be available

The video features provide high level of fidelity for precision localization that is not possible with a head-mounted IMU system alone. The localization method is based on an error-state Kalman filter algorithm using both relative (local) measurements obtained from image based motion estimation through visual odometry, and global measurements as a result of landmark/ object matching through the pre-built visual landmark database. Exploiting the multiple-sensor data provides several layers of robustness to the navigation system. Figure 3 shows the navigation component of the AR system in operation for a MOUT dismount training task, where Avatars are overlaid on the trainees display.

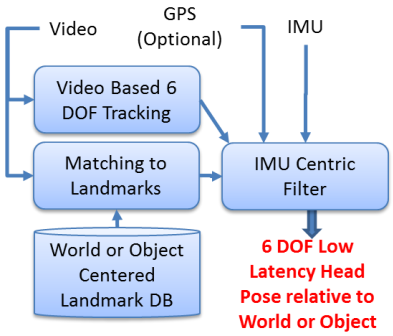

Figure 3: User Tracking and Geo-Position Components.

Automatic Speech Recognition

The Automatic Speech Recognition (ASR) module converts speech to text and can be customized to a specific domain by developing the language and acoustic models. ASR is based on developing models for a large-vocabulary continuous-speech recognition (LVCSR) system that integrates a hierarchy of information at linguistic, phonetic, and acoustic levels. ASR supports natural, spontaneous speech interactions driven by user needs and intents. This capability contrasts with most IVR systems where the system directs the dialogue, and the user is constrained to a maze of questions and limited answers. In addition, ASR can also support speaker-independent spontaneous speech when the topic of the conversation is bounded to a specific domain.

The ASR system used here has a much larger vocabulary than the previous commercial IVR systems, and is based on statistical methods that learn from large amounts of data. ASR uses a sophisticated statistical model to characterize the acoustic realization of the sounds of a language, and to accurately discriminate among a very large set of words (this statistical model is known as the "acoustic model"). ASR also uses a second statistical model to characterize the probabilities of how words can be combined with each other. This second model is referred to as the "language model." More technically, the language model specifies the prior probability of word sequences based on the use of N-gram probabilities. For the AR-Mentor system to perform optimally, the training data must be as representative as possible of the actual data that would be seen in the real system operation. This in-domain data is necessary in addition to publicly available, out-of-domain data to complement the training of the statistical models.

Natural Language Understanding

The Natural Language Understanding (NLU) module is responsible for transforming the user's utterance in natural language, using speech input, into a machine-readable semantic representation of the user's goal. The NLU module can be divided into two sub-modules:

- Event/intent classification: Determine the user goal in a given utterance
- Argument extraction: Determine the set of arguments associated with the user goal

The most challenging issue in NLU is the issue of meaning representation (i.e., representation of intents and their arguments). Human language expresses meaning through various surface forms (e.g., prosody, lexical choice, and syntax), and the same meaning can be expressed in many different surface forms. These aspects are further accentuated in conversational systems, in which the dialogue context plays a significant role in an utterance's meaning. Another aspect that is particularly important for spoken language understanding (SLU) is robustness to noise in the input. Unlike that of text understanding, the input to SLU is noisy because it is the output of a speech recognizer. In addition to this noise, spoken language is rampant with disfluencies, such as filled pauses, false starts, repairs, and edits. Hence, in order to be robust, the SLU architecture needs to cope with the noisy input from the beginning and not as an afterthought. Also, the meaning representation should support robust inference even in the presence of noise. To overcome these challenges, we employ a hybrid understanding strategy in the NLU engine, which uses frame-based semantics, of:

1. High-accuracy, domain-specific rule-based understanding system based on top-down recursive transition network chart parsing, and
2. More generic state-of-the-art statistical intent classification and argument extraction systems based on maximum entropy classification.

Task/Mission Understanding

The Task/ Mission Understanding component is responsible for recognizing/interpreting the user's goal in a given state/context. The scene and language understanding components described above provide partial information about what the user is trying to do at a given time but usually individual components do not have access to all the information required to determine the user's goal. The primary objective of the task/missing understanding component is to merge pieces of information coming from different components, such as scene and language understanding in this case, as well as information coming from previous interactions, i.e., context/state information.

For example, the user might look at a particular object and say "where do I put this?" The scene understanding component will identify the location of objects in the scene and direction that the user is looking at (e.g., a screwdriver), and the language understanding component will identify that the user is asking a question to locate the new position of an object but neither component has a complete understanding of user's real goal. By merging information generated by individual components, the system will determine that the user is "asking a question to locate the new position of a specific screwdriver." Furthermore, most of the time, it is not enough to understand what the user just said in the last utterance but it's also important to interpret that utterance in a given context. In the running example, depending on the task the user is trying to complete, the question in the utterance might be referring to a "location for storing the screwdriver" or a "location for inserting the screwdriver into another object." The task/missing understanding component in this application merges three different semantic frames representing three different sources of information at any given time:

1. Semantic frame representing the scene (from scene understanding)
2. Semantic frame extracted from the last user utterance (from language understanding)
3. Semantic frame that represents the overall user goal up to that point (from prior interactions).

The task/missing understanding component can also utilize useful information about the user's history and characteristics to augment the context information, which could enable adapting and customizing user interaction.

Merging of these three pieces of information is accomplished using a hybrid approach that consists of:

1. A domain-independent unification mechanism that relies on an ontology structure that represents the events/intents in the domain
2. Task-specific workflows using SRI's VPA workflow execution engine.

Virtual Personal Assistant Reasoning

The VPA Reasoning module's function is to determine what the AR-Mentor should do next. Specifically, the VPA Reasoning module takes the detailed representation of the user's current state and goal, as determined by the Task/Mission Understanding component, and produces a representation of an appropriate system response, where

the response may be dialog, UI displays, or some combination of the two. This task requires detailed domain knowledge to ensure that the AR-Mentor responds correctly and takes appropriate action from a domain perspective, and that these responses and actions instill *trust* in the user. Reasoning must calculate the AR-Mentor's next response or action using a variety of diverse sources: detailed knowledge of the domain's procedures and preferred styles of interaction; known information about the user, including their level of expertise in the domain; and the status of the context of the dialog with the user this far.

The detailed architecture of the VPA Reasoning module is shown in Figure 4. The architecture facilitates the acquisition of multifaceted domain knowledge (the left third of Figure 4) designed to drive user-system dialogs and interactions covering a wide variety of topics within the domain. This knowledge is then compiled by an engine (the middle third of Figure 4) into machine-interpretable workflows along with (if necessary) a set of methods that interact with domain back-end systems—retrieving information from legacy databases, etc. Then at run time (the right third of Figure 4), the VPA Reasoner uses those compiled workflows to interpret User Intents received from the Understanding module and determine the next step for the system to take. This step is represented as an AR-Mentor intent, and may encode dialog for the AR-Mentor to utter, actions or changes within the UI, both of those, or even neither of those (i.e., take no action).

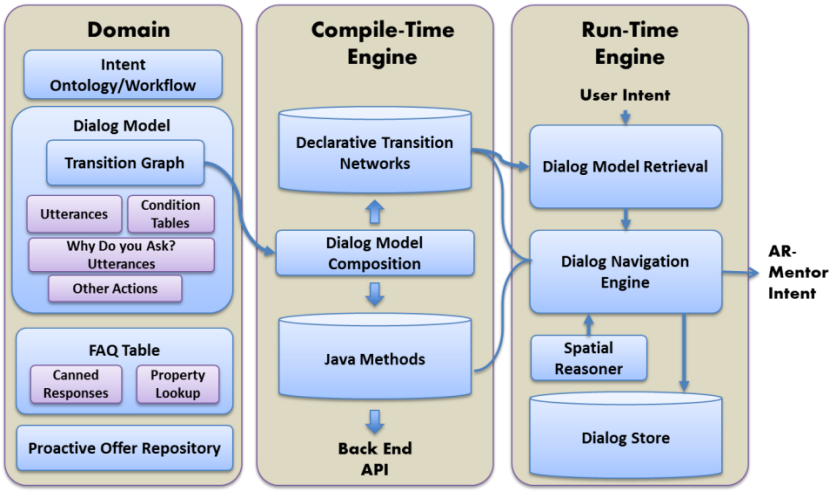

Figure 4: VPA Reasoning architecture.

Content (Speech and Animation) Generation

Content generation module takes input from the reasoning system and creates the visual overlays and corresponding natural language dialog that is presented to the user.

Augmented Reality 3D Animation Generation Module

The AR 3D Animation Generation module uses computed head poses to accurately render animations and instructions on the AR Goggles so that the rendered objects and effects appear as if they are part of the scene. The low-lag realistic overlays that match precisely with the real-world is one of the hardest challenges of the AR-Mentor system.

The critical ingredients for produces good results are provided by the scene understanding module. This module relies on the localization engine in the Scene-Understanding module to obtain accurate head pose. It is important for the pose generate to account for delays in the video processing and rendering latencies if the overlays are to correctly appear in the world. As such the animation generation module asks the localization module to predict a pose just-in-time for rendering. On such request the localization modules Kalman Filter is able to exploit the high-rate IMU in the system to very accurately predict the location and orientation of the head. In previous systems that we have evaluated we were able to predict the motions of the head to within 5-10 milliseconds.

The second critical aspect to rendering is handling occlusions. This module is able to work with dynamic depth maps in its rendering pipeline. The dynamic depth that is obtained from the scene understanding module is fused with information from CAD-models (for the scene or objects) that are available to create consistent occlusion masks for the rendering. This ensures correct 3D layering between the rendered objects against the real-world. In addition to the core 3D animation module we have incorporated a labeling system for labeling objects in the scene and organizing these labels on the rendered view.

In the back-end the animation system relies upon a well-organized pre-authored domain specific content to enable intuitive instructions. The authored content is organized hierarchically and incorporated within the VPA Reasoner

logic to ensure intuitive triggering of these scripts. To support this we have incorporated an intuitive scripting language that can be used by the VPA Reasoner. Based on these higher level instructions rendering engine will sequence through lower-level set of animations and visualizations with intuitive transitions.

Natural Language Generation and Text-to-Speech Module

The output generation module receives system actions from the reasoning component and converts them into different forms of action representations, such as text, speech, domain specific actions, and UI manipulations, as appropriate for the user and the environment.

The Natural Language Generation **(**NLG) module employs hierarchical output templates with fixed and optionally variable portions that are generated on the fly using linguistic tools to generate system responses in a given interaction with the user. Each action generated by the VPA Reasoning component has an associated prompt template, and the system chooses the most appropriate response by synthesizing the variable portion of the response. Finally, we use a commercially available Text To Speech product from NeoSpeech for verbal response.

AR-Mentor Hardware

Figure 5 shows our customized 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 HMD 1). The stereo cameras are arranged vertically for minimal intrusion to the user, and the images (640x480) are captured at 15fps and the IMU unit operates at 100HZ. The stereo cameras and the IMU unit 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 navigation unit can be transformed by the system to know where to insert synthetic objects in the HMD. The current sensor rig weights around 1 lb. The processing was divided between a user-worn computer and a standalone remote laptop running as a server. The non-time-critical processing components such as VPA and ASR ran on the standalone remote laptop server and the time-critical processing tasks including user head localization and HMD rendering ran on an Apple Mac-Mini computer that the user is wearing. The computers communicate wirelessly via a WIFI router. The Apple Mac Mini is quite compact and equipped with 2.3GHz Quad-Core Intel Core i-7 CPU.

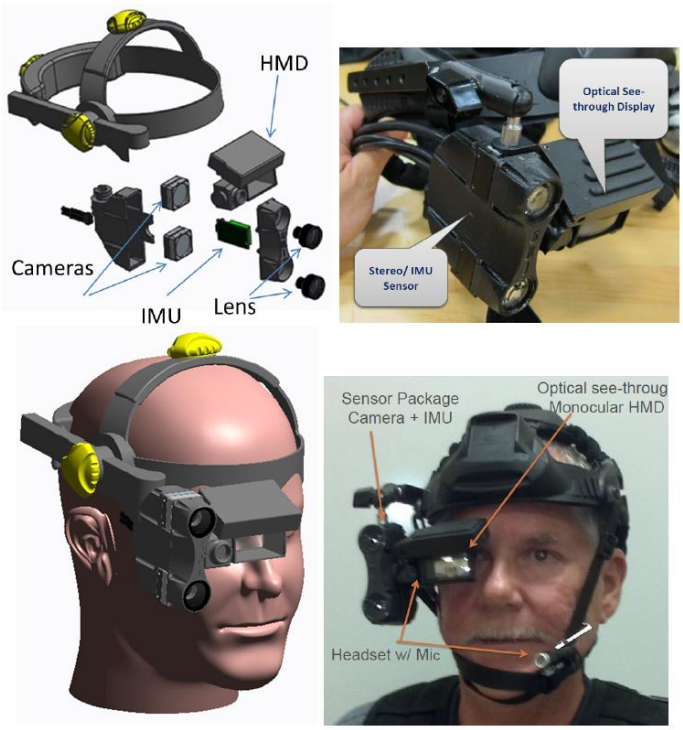

Figure 5: 3D Rendering engine incorporates predicted pose and occlusion reasoning from the scene understanding module to accurately render 3D content with low latency.

EXPERIMENTAL RESULTS

We developed a pilot application of using AR-Mentor to help student mechanics learn and perform an advanced maintenance procedure on a military vehicle. There are 33 steps written in 11 pages in the technical manual, and it usually takes approximately 40 minutes for an experienced training vehicle mechanics to perform. The task requires the mechanic to move to four different locations to perform: inside the turret, on the top of the vehicle, on the left side of the vehicle and inside the cargo hatch.

It's a complex advanced procedure and involves the use of a new tool (the bubble level). The school 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 get it right, and it is a very high burden for the instructor. The AR-Mentor is designed to help relieve this burden.

Figure 6 shows a set of selected example images with virtual insertions for step 6, which consists of 5 sub-steps that instruct the student how to zero a turret level on a mounting plate step by step. Figure 7 shows a set of selected example images with virtual insertions for step 10, which will instruct the student how to use a ratchet wrench to remove the shield from housing by removing four screws in order. Two captured videos of both steps with audio are provided in the supplementary materials. Note these videos were recorded for demonstration in video mode.

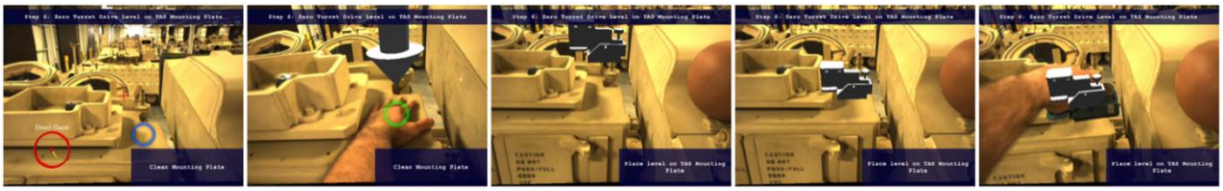

Figure 6: A set of selected example images of Step 6 with virtual insertions (tools, parts and texts).

Figure 7: A set of selected example images of Step 10 with virtual insertions (tools, parts and texts).

Evaluation

A formative, descriptive "Level 1" (Kirkpatrick, 1975) evaluation was used because of the early-stage nature of the prototype and the constraints of time and access to mechanics of the testing context. The evaluation examined the usability of the system and how the technology affected the learning experience by comparing three different learning conditions (AR-Mentor only, Instructor/technical manual, and technical manual only). We included the technical manual only condition for two reasons: (1) To support comparability of our study to past AR research, which almost exclusively compared the technology to unassisted learning, and (2) to provide a baseline of actual moments when learners needed help to learn the task. Pilot testing and observations of standard schoolhouse instruction indicated that instructors used intrusive guidance with student mechanics, meaning they monitored the novices' performance closely and intervened before novices requested assistance. This instructional approach can mask actual learner needs. All participants were recruited in accordance with human subjects regulations and procedures. The novice student mechanics had never been taught this specific procedure before and were divided between high- and low-scorers on a multi-aptitude test. We engaged pairs of learners in conducting the 33-step training vehicle maintenance task, with one learner conducting most of the hands-on work and the other helping. In the instructor-guided and technical manual conditions, the "helper" reads the technical manual; in the AR-Mentor condition, the helper served as a "second set of hands." In the AR-Mentor condition only, learners switched roles and repeated the procedure to provide more input on the usability of the system and show how these teams improved the second time. In the other two learning conditions, the pairs went through the procedure only once to provide comparative performance data. Instructors were on-hand in all three conditions to intervene if students made an error that would impede progress, but they were asked to avoid intervening in the AR-Mentor and technical manual only conditions. By contrast, they were asked to provide guidance as they normally would in the instructor condition.

To measure usability, we administered a questionnaire and interview about satisfaction to the novice mechanics in the AR-Mentor condition. To measure learning experience, we developed two protocols for all three learning conditions: An observational protocol and a learning experience questionnaire. The observation protocol recorded the following performance data: Time to complete each of three sub-phases of the adjustment, total task completion time, number of total errors, and number of total instances of help (either sought or intrusively provided by the instructor). It also included verbal concept checks at the end of the preparation, testing, and reassembly phases to see if learners understood electrical circuitry. The learning experience questionnaire focused on learners' perceived

mental demand of the overall task and of selected steps that varied in complexity—some that required multiple sub-steps of preparation and some that were straightforward.

We employed a descriptive and quasi-quantitative analysis (Stake, 1995). For the usability sub-study, we examined the interview results and the ratings of the four novices in the AR-Mentor condition, both individually and in aggregate. For the learning experience sub-study, we tallied phase completion times and created a comparable rating of "observed learner need" across the three conditions by tallying the number of errors made and number of either student- or instructor-initiated requests for help. We analyzed the learning experience questionnaires by generating descriptive statistics of mental demand ratings, overall and per selected steps. In a second-level analysis, we classified 8 steps as "complex" and the remaining steps as "simple." Finally, we compared the AR-Mentor representations for these simple and complex steps by counting the number of AR visual representations and utterances and estimating the mental transformations required to map them to the work task. Each pair of mechanics was anonymized and assigned a label, as follows: P1.1 for the first user and P1.2 for the second user in Pair 1, etc.

In the satisfaction questionnaires and interviews for the AR-Mentor learners, four mechanics reported liking the AR-Mentor verbal and visual representations (see Table 1). As shown in Table 1, novice student mechanics gave high ratings to the overall usability for AR-Mentor. However, interviews indicated they wanted to shorten the verbal dialog as needed and preferred a smaller form factor for the hardware to use the device in tight spaces. The concept checks and behavioral observations showed a bimodal performance in all three conditions. Mechanics overall performed worse on concept checks about electrical circuitry on complex steps, which involved positioning, setting up, and reading a testing device: Only 2 out of 8 mechanics got the concept questions correct on these steps. On simple steps, such as adjusting a screw, 7 out of 8 mechanics got the concept check correct.

Table 1. Novice soldiers' perceptions of overall usability of the AR-Mentor

Overall Usability Performance metrics for AR-Mentor. Ratings are in scale of 1-5 (Low to High)	**Soldier ID**				**Average Rating**
	P1.1	**P1.2**	**P2.2**	**P2.3**	
How well could you understand what you needed to do to adjust the Bradley?	5	2	4	5	4.00
How consistently did the AR-Mentor audio and visual instructions occur when you needed them to occur?	4	4	5	4	4.25
How consistently did you know which steps required you to coordinate with a helper and which required only you?	4	3	5	5	4.25
Overall					4.17

Table 2 illustrates the counts of performance metrics, and it shows that the AR-Mentor condition took slightly longer and involved a few more errors than the instructor condition overall; novice student mechanics took longer and made more errors in the manual-only condition. The data on help-seeking and total instances of instructor intervention indicate that the AR-Mentor required much less guidance from the instructor and that that instructors were intervening more often than learners expressed a need, as compared to the baseline manual-only condition (see Table 2). The novices working with the AR-Mentor were observed to be engaged in much more independent, collaborative troubleshooting than those in the instructor condition.

Table 2. Comparison of trainee error, help-seeking, instructor guidance, and time per learning conditions.

Learning Condition	**Total Errors Mean**	**Total Help Seeking Mean**	**Total Instructor Guidance Mean**	**Average Total Time:hr:min:sec**
AR-Mentor	2.75	7.50	2.00	1:11:00
Instructor+Manual	1.50	8.00	46.50	0:55:30
Manual only	8.00	25.00	22.50	2:15:00

Additional analysis focused on the bimodal results on complex and simple steps to understand why the AR-Mentor representations were not working as effectively on complex steps. Analysis indicated that AR-Mentor included more dense representations around complex steps (M: 6 visuals; 13 utterances) than simple steps (M: 1.7 visuals; 3.7 utterances). Analysis of the mental transformations required around an introductory visual representation used for

the complex steps indicated learners needed to engage in as many as 39 instances of spatial rotation and inference to map it to reality as compared to 2-4 on average. Analysis of the duration of the introductory segment of verbal dialog around the complex steps indicated that learners had to absorb 47 seconds of explanation as compared to 1 to 5 seconds on average.

We concluded that the AR-Mentor introductory representations for complex steps were so dense as to impose extraneous mental demand on the learners. Revisions are suggested to break up the density of the AR-Mentor visual and verbal representations from a "full introduction" approach to a "dosed sub-step and post-action review" approach for the complex steps.

CONCLUSION AND FUTURE WORK

In this paper, we described a highly flexible, mobile, automated mentoring and training system, AR-Mentor, tailored to the needs of individual users. The system will act 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 will interact with the user in natural spoken language and through visual (AR) indicators in the scene. It will provide training in both the operation and maintenance of the user's equipment. This proposed AR-Mentor prototype is the first mentoring system that uniquely combines AR, visual and NLU and reasoning and VPA technologies together. The formative evaluation supports its basic usability and indicates that it offers a learning experience that permits learners to focus on their needs with an acceptable levels of mental demand overall. In the future, we will improve the AR-Mentor in a number of directions. 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. Second, we will reduce the weight of both computer and batteries by using mobile processors. Finally, on the VPA side, we will focus on simplifying the authoring process for procedural tasks and injecting more robust diagnostic reasoning into VPA. Additionally, the formative evaluation indicated some opportunity to differentiate the design of graphic and dialog representations between simple and complex steps in specific procedural areas.

ACKNOWLEDGEMENTS

The material presented in this paper is based upon research 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

P. Chandler. Dynamic visualizations and hypermedia: Beyond the "Wow" factor. Computers in Human Behavior, 25(2):389–391, 2009.

M. Hegarty, S. Kriz, and C. Cate. The roles of mental animation and external animation in understanding mechanical systems. Cognition and Instruction, 21:325–360, 2003.

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.

Y.-D. Kim and I.-Y. Moon. E-training content delivery networking system for augmented reality car miantenance training application. International Journal of Multimedia and Ubiquitous Engineering, 8(2), 2013.

R. Kumar, S. Samarasekera, A. Chaudhry, Z. Zhu, H. P. Chiu, T. Oskiper, R. Villamil, V. Branzoi, R. T. Hadsell, E. R. Pursel, F. Dean, and P. Garrity. Implementation of an augmented reality system for training dismounted warfighters. Interservice/Industry Training, Simulation, & Education Conference (I/ITSEC), Orlando, FL, 2012.

K. Lee. Augmented reality in education and training. TechTrends, 56:13–21, 2012.

R. E. Mayer and R. Moreno. Aids to computer-based multimedia learning. Learning & instruction, 12(1):107–119, 2002.

T. Oskiper, M. Sizintsev, V. Branzoi, S. Samarasekera, and R. Kumar. Augmented reality binoculars. IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2013.

S. Webel, 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.

S. Yantis and J. Jonides. Abrupt visual onsets and selective attention: voluntary versus automatic allocation. Journal of Experimental Psychology: Human perception and performance, 16(1), 1990.

D. L. Kirkpatrick. (Ed.). Evaluating training programs. Tata McGraw-Hill Education, 1975.

R. Stake. The art of case study research. Thousand Oaks, CA: Sage, 1995.