Towards a Warfighter's Associate: Eliminating the Operator Control Unit

H.R. Everett,[a] E.B. Pacis,[a] G. Kogut,[a] N. Farrington,[a] S. Khurana[b]

[a] Space and Naval Warfare Systems Center, San Diego (SSC San Diego)
[b] University of Southern California (USC)

ABSTRACT

In addition to the challenges of equipping a mobile robot with the appropriate sensors, actuators, and processing electronics necessary to perform some useful function, there coexists the equally important challenge of effectively controlling the system's desired actions. This need is particularly critical if the intent is to operate in conjunction with human forces in a military application, as any low-level distractions can seriously reduce a warfighter's chances of survival in hostile environments. Historically there can be seen a definitive trend towards making the robot smarter in order to reduce the control burden on the operator, and while much progress has been made in laboratory prototypes, all equipment deployed in theatre to date has been strictly teleoperated.

There exists a definite tradeoff between the value added by the robot, in terms of how it contributes to the performance of the mission, and the loss of effectiveness associated with the operator control unit. From a command-and-control perspective, the ultimate goal would be to eliminate the need for a separate robot controller altogether, since it represents an unwanted burden and potential liability from the operator's perspective. This paper introduces the long-term concept of a supervised autonomous *Warfighter's Associate,* which employs a natural-language interface for communication with (and oversight by) its human counterpart. More realistic near-term solutions to achieve intermediate success are then presented, along with actual results to date. The primary application discussed is military, but the concept also applies to law enforcement, space exploration, and search-and-rescue scenarios.

Keywords: robotics, autonomous systems, augmented reality, natural language understanding, sign interpretation, speech recognition, simultaneous localization and mapping, world modeling, collision avoidance, target acquisition, machine vision**.**

1. BACKGROUND

The *ROBART* series of autonomous research prototypes has served in developing the component technologies needed in support of the Mobile Detection Assessment Response System (MDARS) robotic security program.[1] While *ROBART I* (1980-1982) could only detect a potential intruder,[2] *ROBART II* (1982-1992) could both detect and assess, thereby increasing its sensitivity (i.e., probability of detection), with a corresponding reduction in nuisance alarms.[3] Other research thrusts included implementation of an absolute world model, automated localization techniques to null out accumulated dead-reckoning errors, and reflexive (sensor-assisted) teleoperated control concepts for guarded motion.

Fig 1. *ROBART III,* the development platform for the *Warfighter's Associate*.

The third-generation prototype, *ROBART III* (1993-) was originally intended to demonstrate the feasibility of automated response, using a pneumatically powered six-barrel Gatling-style weapon that fires simulated tranquilizer darts or rubber bullets (Figure 1). Early work extended the concepts of reflexive teleoperation into the realm of coordinated weapons control (i.e., sensor-aided control of mobility, camera, and weapon functions). Starting in FY-03, the navigation and collision avoidance schemes are being significantly enhanced through technology transfer of improved algorithms developed under DARPA's Tactical Mobile Robot (TMR) and Mobile Autonomous Robot Software (MARS) programs.[4] Appropriate hardware upgrades (to include a MicroStrain gyro-

stabilized compass, KVH fiber-optic rate gyro, SICK scanning laser rangefinder, Visual Stone 360-degree omni-cam, and Canon pan-tilt-zoom camera) have also been made to support the more sophisticated navigation, collision avoidance, mapping, and surveillance schemes. For these and other reasons, *ROBART III* was selected as the optimal laboratory development platform for investigating ultimate feasibility of the *Warfighter's Associate* concept.

2. INTRODUCTION

Recent and ongoing military actions in Afghanistan and Iraq marked the first time robotic systems played a meaningful role during actual combat operations, supporting cave and bunker reconnaissance, chemical and radiological detection, and explosive ordnance disposal (EOD) missions. EOD units from all four military services are currently using a variety of teleoperated systems on missions ranging from scouting unsecured bunkers, buildings, or caves, to neutralizing improvised explosive devices (IED), and there is increasing demand for more robots with even more capabilities. Accordingly, the focus of SSC San Diego's spiral development program is to improve the autonomous functionality of this baseline hardware and incorporate additional application payload modules, so as to provide increased utility with less of a control burden imposed upon the operator.

Extrapolating out to some point in the future, one can almost envision a sophisticated robotic system, ultimately perhaps even anthropomorphic in nature, intended to routinely accompany a warfighter in the execution of his or her mission. This futuristic *Warfighter's Associate* would be specifically designed and equipped to exploit its complimentary robotic strengths, thus enabling a very synergistic teaming of human and machine capabilities. A good analogy here can be seen in the pairing of police officers and their canine partners in both military and civilian law-enforcement applications; each player has some rather unique talents that enable the resulting K-9 team to achieve impressive results.

Relative to machines, humans have a number of well known disadvantages: tend to tire easily, need sleep, can become bored or otherwise distracted, and are susceptible to disease. The human body is extremely frail, and accommodating its life-support and creature-comfort needs on a battlefield can be very expensive, particularly when hostile forces are actively trying to exploit these vulnerabilities. Nevertheless, humans are remarkably perceptive, extremely adaptive, and very flexible, capable of quickly reasoning out complex solutions to unexpected and/or changing conditions. These attributes make humans essentially indispensable and suggest that the role of the robot will for the most part remain subservient in a supervised-autonomous capacity, as opposed to fully autonomous.

On the other hand, the computers that serve as a robot's distributed brain are less adept at these human-suited tasks but excel at such mundane and often tedious things as storing images or map representations, calculating precise absolute location in real time, sorting large amounts of data to detect patterns or anomalies, and network communications. When interfaced with appropriate sensors, they support non-contact range measurement, night vision, chemical and radiological detection, as well as location of landmines and IEDs.

Obviously the envisioned *Warfighter's Associate* would handle any high-risk tasks that involved exposure to hostile conditions, and potentially could physically evacuate an injured human if the need arose. The system could provide on-site interpretation of foreign languages, detailed repair and maintenance instructions for organic equipment, even awareness of medical procedures to augment the training and knowledge of medics in the event of human casualties in the field. With Internet access, the robot would furthermore be able to research emergent topics of interest in near-real-time, providing an essentially limitless knowledgebase of valuable information at the human's request.

3. TECHNICAL CHALLENGES

While the concept of a humanoid *Warfighter's Associate* is highly ambitious for the near term, more practical embodiments are arguably feasible, with substantial improvements likely to come along later as the technology continues to evolve. This section presents a quick review of where technology areas currently stand, along with reasonable projections for the future.

3.1 Mobility

Ultimately it would be advantageous to have an anthropomorphic (or at least legged) configuration for improved mobility in rugged terrain or battle-damaged structures. After all, people and animals use their legs to achieve amazing

agility and consequently can go places where no tracked or wheeled vehicle could venture. On the other hand, some of the smaller man-portable robots (i.e., the iRobot *PackBot*) can get into tight spaces where a humanoid robot would have serious difficulty. DARPA's innovative *RHex* configuration (Figure 2) is both legged and small, and may ultimately prove to be advantageous. Despite the demonstrated dexterity of the very impressive humanoids recently introduced by Honda, Sony, and others, these prototype units are still unable to adaptively cope with real-world surfaces for which their movements were not specifically pre-taught. Truly adaptive legged locomotion is making great strides at places such as Boston Dynamics, Tokyo Institute of Technology, and Waseda University in Japan, but still has a ways to go. So while a few niche (i.e., rough-terrain) applications call for legs, it is expected that wheeled and/or tracked vehicles will remain the mobility solutions of choice for the near-term, especially reconfigurable versions that automatically adapt to their immediate environment and tasking.

Fig 2. DARPA's *RHex* prototype employs six dynamically controlled rotating "legs" for enhanced mobility in rough terrain.

3.2 Navigation

For purposes of this discussion, the term navigation covers those subtasks required for the robot to figure out where it is, plan a path to where it needs to go, and then get there without running into anything. Relatively speaking, these technical challenges (i.e., localization, path planning, collision avoidance) have for the most part been reasonably solved. The MDARS program,[1] for example, has been operating autonomous robotic security systems in both indoor and outdoor environments for a number of years. GPS has effectively addressed the localization problem in outdoor environments, and the DARPA TMR and MARS programs have significantly advanced the state of the art for indoor localization and mapping. A number of fairly adept collision-avoidance schemes have been produced (with DARPA again making major contributions through the TMR, MARS, and PerceptOr programs), and the biggest limitation remaining here is availability of a suitable (i.e., small, light-weight, low-power) sensor suite to support these algorithms on the smaller man-portable robots. SSC San Diego is currently evaluating a miniature stereo system developed under TMR by the Jet Propulsion Lab (JPL), and conducting a market survey for candidate scanning laser rangefinders that may help meet this need.

3.3 Power

The perception, computational, and actuation schemes required for a supervised autonomous robot that could perform as a *Warfighter's Associate* will collectively require some considerable power, and providing a reliable, safe, easily renewable energy source that can handle these needs over extended periods of time (i.e., roughly human equivalent) remains a big problem. Conventional batteries on current man-portable systems last only about four hours, and these systems are nowhere near as complex or power hungry. Solar power has been effectively employed for applications in space, such as the Mars Rovers built by JPL, where speed and endurance have been sacrificed for longevity, but is ill suited to most military applications. Fuel cells offer some near-term promise, particularly those using alcohol as opposed to hydrogen as a fuel, in that the latter cannot be transported on military aircraft due to safety restrictions that ban the requisite high-pressure (2000 psi) containers. Accordingly, a major technological breakthrough is needed here for the long term, and until then the needs will most likely be met by hybrid fossil-fuel/electric systems, with the attendant tradeoffs in capability.

3.4 Command and Control

In the beginning, there was almost always a one-to-one correspondence between the mobile robot and some dedicated host computer that served as the remote OCU. (One notable exception was the completely autonomous *ROBART I*, which had no OCU at all.) The MDARS command and control architecture eliminated this direct association with a dedicated controller, allowing multiple robot control, including robots of different types.[1] Other efforts facilitated hand-off from one controller to another, allowing multiple operators to talk to the same robot. SSC San Diego's *Multi-Robot Operator Control Unit (MOCU)* shown in Figure 3, for example, provides a standardized controller with plug-and-play I/O capability,[5] based on the Joint Architecture for Unmanned Systems (JAUS) mandated by the Office of the Secretary of Defense (OSD).

From a command-and-control perspective, however, the ultimate goal in a tactical environment would be to eliminate the need for a separate robotic controller altogether (at least at the organic level), since it represents an unwanted burden and potential liability for the operator. Today's warfighters have enough equipment to carry as is, and anything that needlessly distracts them with low-level details can seriously reduce their chances of survival in hostile environments. Currently there is a tradeoff between the value added by the robot (i.e., in terms of how it contributes to the performance of the mission), and the additional burden imposed by the OCU (i.e., how it interferes with the operator's ability to perform and perhaps even survive).

Fig 3. The *MOCU* controller, shown here in a back-packable configuration, can control a variety of SSC San Diego's unmanned air, ground, and surface vehicles.

There are development programs underway to equip our troops on the ground with secure digital communication devices for bidirectional voice, video, and map displays.[6] From a situational-awareness perspective, there is an obvious tactical advantage to be gained by one soldier relaying secure video from his or her vantage point to other members of the squad. If a robotic system could seamlessly take the place of that potentially exposed and vulnerable human reconnaissance source, impervious to chemical and biological agents, and equipped with more effective surveillance sensors, so much the better. The natural objective here would be for the *Warfighter's Associate* to use this same communication mechanism and interact no differently than a human.

4. NATURAL LANGUAGE UNDERSTANDING

Accordingly, SSC San Diego is pursuing a natural-language interface that would allow the *Warfighter's Associate* to be given fairly unstructured verbal direction, no different from the procedures used to instruct a human to perform the same task. If this concept seems a bit too futuristic, the far end of the spectrum is probably better represented by efforts underway at Duke University (and other organizations) to directly control a robot using human thoughts.[7] Researchers under the direction of Miguel Nicolelis, Co-Director of Duke's Center for Neuroengineering, have demonstrated rudimentary control of a robotic manipulator based on decoded neural activity in the brain of a macaque monkey. The current setup requires electrodes to be implanted in the monkey's brain, which obviously is a little too intrusive for humans, but proponents predict non-invasive (i.e., CT-scan) helmets will one day be able to collect the same descriptive neural patterns without need for surgical implants. So in comparison, the concept of reliable interactive speech between man and machine looks considerably less challenging for the near term.

The *Warfighter's Associate* concept envisions a bi-directional natural-language interface that needs no robot-specific hardware, which means the robot must be able to both generate speech output as well as understand speech input. The first of these requirements, speech synthesis, is a very mature technology. *ROBART I*, for example, could vocalize 256 words back in 1981, and today's text-to-speech algorithms are very robust, with essentially unlimited vocabularies. Understanding speech, however, is significantly more complicated and involves two fundamental issues: 1) recognizing the spoken words, then, 2) parsing the resultant text.

4.1 Recognizing Words

Although reliable speech-to-text algorithms have successfully found their way into a number of commercial voice-recognition applications, most exploit a fairly high signal-to-noise ratio with respect to the incoming audio stream. That is to say, the user is typically talking directly into a microphone, such as a boom mike or telephone mouthpiece, and ambient noise conditions are minimal. On the battlefield, a number of problems can arise, for a war zone is an inherently noisy environment. In addition, the strain of combat can easily alter a pre-taught voice signature, in that humans tend to talk louder, faster, and with an increase in pitch when under stress associated with noise and danger. Finally, there are times when absolute silence must be maintained for purposes of stealth, and talking is not allowed. Bone-conduction headsets and microphones, acoustically coupled via the bone structure of the scull, have been shown

Fig 4. *Jawbone* active noise-canceling headset by Aliph.

to offer significantly improved performance under some of these types of conditions. *Jawbone* (Figure 4), a more recent introduction from Aliph (Brisbane, CA) targeting the cell-phone market, goes a step further by employing two microphones and a DSP to subtract background noise from the desired vocal input, using an innovative bone-conduction sensor to determine when the user is actually speaking.

It's worth noting that Jakks Pacific, Inc. recently upgraded its popular *R.A.D.* toy robot (formerly distributed by Toymax) to employ a voice-recognition interface in time for the 2003 Christmas season (Figure 5). The system recognizes 50 speaker-independent commands, has a 500-word speech-synthesis capability, and is equipped with a three-shot missile launcher, all for $39.99! Recognition reliability is fairly robust, even at distances of around ten feet from the head-mounted microphone, as long as ambient noise is low. While this simplistic (but impressive!) remote-control toy lacks the appropriate sensors and processing electronics to support intelligent behavior, it clearly demonstrates the growing feasibility of interactive voice control. The point to note is the real utility of this concept goes up substantially as the robot gets more intelligent: the control paradigm shifts from low-level teleoperation to high-level supervision, and the burden on the operator decreases accordingly.

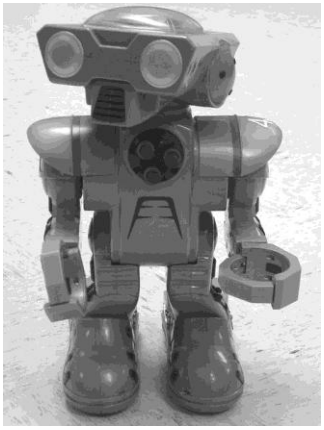

Fig 5. *R.A.D. 4.0* incorporates a 50-word voice-input command set. (Note similarity of the head to that of *ROBART III*.)

4.2 Parsing Text

Once the recognition algorithm has converted incoming speech to a text string, the parsing algorithm must next transform the text string into a suitable command. This step is essentially bypassed in the more simplistic schemes, such as the example toy application above, by using a layered menu of all possible commands, along with a very structured command set. There is a predefined mapping of commands at each layer, and only one word is considered at a time, so there is no need to truly "parse" any text. Each word in the command phrase simply determines a conditional branch through a predefined tree structure, which in turn identifies which subgroup of words should then be examined for a match to the next word of the spoken command. In the previous case of the toy, the first word of a three-word command format is always the name of the robot, which alerts it to expect two more words that collectively describe the desired action.

The second word tells the system which subgroup of final words to expect and also the type of requested command, such as "*move.*" The final word represents a parameter that further defines how the command is to be executed. For a *move* command, typical parameters include *left*, *right*, *forward*, or *backward*. Thus the spoken command, "RAD, move, left," would result in a controlled branching within the tree structure of predefined possibilities, ending with the one that corresponded to a left turn. An early speech-recognition system evaluated on *ROBART II* in the mid eighties employed this same approach, with 16 groups of 16 words each. In addition to simplistic parsing, recognition reliability is also significantly enhanced under this scheme, because each incoming word is compared against only 16 possibilities (in this particular case) in a multiple-choice format.

Parsing truly unstructured text is much more complicated, but some very amazing examples of success have been around for quite some time. One of the most well known is the artificial intelligence personality *"Eliza,"* created in the sixties by Joseph Weizenbaum of MIT. *Eliza* was a natural-language processing system programmed to emulate a Rogerian psychiatrist conducting an exploratory interview with a new patient.[8] The "patient" would converse with the program via keyboard input, with *Eliza* restructuring the patient's single-sentence statements into seemingly related intelligent responses to keep the conversation moving along. The results were pretty impressive, especially for the time, and did not require enormous computational resources to execute. In fact, a version of the program available in the mid-eighties was briefly installed for evaluation on *ROBART II*, running on a 1-mHz 6502 processor with only 32 kilobytes of RAM.

As previously implied, however, humans are much more adept than computers at interpreting unstructured speech under dynamic and distracting battlefield scenarios. But even human-to-human communication in such circumstances is purposely structured, with established radio procedures and reporting formats aimed at minimizing miscommunication. A typical example would be the use of the terms "affirmative" and "negative" in place of their somewhat harder to distinguish single-syllable "yes" and "no" equivalents. Similarly, rather than indicating a question through voice

inflection, as is often done in normal conversation, the procedure calls for semi-formal structure as follows: "Interrogative your current position, over?"

So in reality, there is no inherent *Warfighter's Associate* requirement to use truly unstructured communication in the purist sense, which makes the problem much more manageable. And unlike some of the more ambitious text-parsing research efforts already underway, this application does not involve reams of unstructured text, but instead needs only interpret short and fairly non-ambiguous single-phrase commands. Revisiting momentarily the earlier police dog analogy, this same philosophy indeed applies: the K-9 handler instructs his four-legged companion with very succinct structured commands, which the latter has been trained to recognize and then promptly execute. There is no need for extended discussion.

An unstructured speech parser was recently developed by Khurana on *ROBART III* to parse unformatted questions/commands/generic statements from both dictated and typed text, in order to extract meaningful words that could elicit the appropriate response. In support of this effort, *ROBART III* was assigned its own e-mail address at SSC San Diego, with the ability to both send and receive messages. This approach not only allowed the parsing algorithms to be developed independently of the effects of speech recognition errors, but also provided a very useful and natural interface for the human-robot team. For example, if the robot needed help transiting a closed doorway and could not detect a local human presence to verbally address, it could broadcast a message requesting assistance to all the occupants of the building. Conversely, it could also receive an e-mail message, perhaps instructing it to go to the sender's office, and from there forward a captured image as a .jpg attachment (Figure 6).

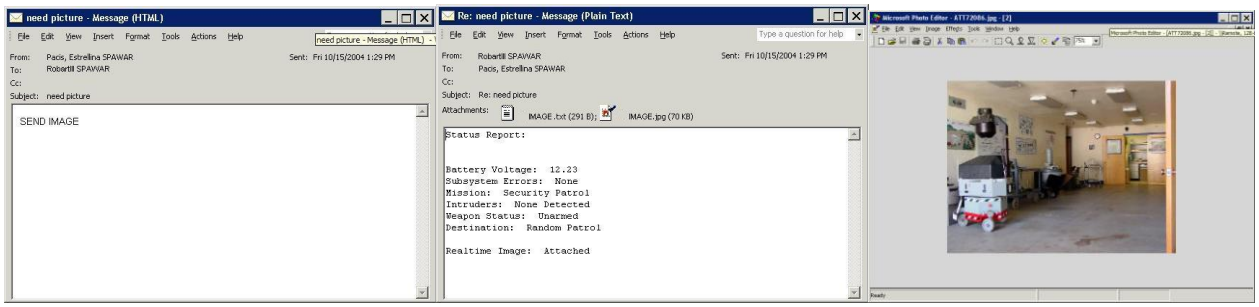

Fig 6. Email request for image and subsequent response with .jpg attachment from *ROBART III* (left to right).

5. EFFECTING THE DESIRED CONTROL

Once the human's spoken commands are converted to text, and the text is parsed to ascertain intent, the problem becomes one of effecting the proper robot response to the given instruction. To facilitate this objective, there needs to be some appropriate frame of reference to which both the human and the robot can unambiguously relate.

5.1 Robo-Centric Reference

The simplest non-ambiguous frame of reference is relative to the robot itself, and such a scheme is clearly adequate for verbalizing low-level motion commands (i.e., turn left, turn right, slow down, stop) during basic or even reflexive teleoperation.[9] A *robo-centric* framework can even be used to initiate sensor-assisted motion primitives (i.e., follow wall on right, enter next opening on left), or to control camera gaze (i.e., pan left, pan right), as was done on *ROBART II*, and later ported over to *ROBART III*. But this simplistic relative approach by itself is insufficient for controlling the more advanced autonomy required for high-level direction, such as the above e-mail example instructing the robot to go to an indicated room and take a picture of a certain area. The *robo-centric* frame of reference in such a context is too restricted in scope, and way too operator intensive in general.

5.2 Vision-Centric Reference

A much more powerful approach would be to use the robot's own camera view as a common frame of reference. For example, suppose the robot has penetrated an underground bunker and is streaming back video that shows an open doorway in the center of the far wall of the room just entered. A human monitoring this video might converse with the robot as follows: "Find the doorway in front of you." The robot would then analyze the current video, looking for

predefined scene attributes that suggest a door frame or opening, highlighting its choice with a graphic overlay. If the robot's vision system locked onto the same doorway the observer had intended, the human would acknowledge as follows: "Affirmative." Or simply say nothing at all.

If for some reason the robot selected the wrong door, however, or a set of scene attributes that was in fact not a door at all, the human would respond differently: "Negative, look to your left." (Or right, as the case may be.) The vision system would shift focus accordingly to the next set of scene attributes that looked like a doorway, again highlight its choice, and so forth. Once the human and the robot were in sync, the human could issue additional voice prompts to influence the robot's further interaction with the identified doorway. One example could be to zoom in on and perhaps even illuminate for better assessment or to enter the doorway and continue searching on the other side.

If the robot were unable to make an appropriate correlation with scene attributes, the human could resort to manually directing the camera gaze and zooming in on the region of interest or interacting directly with the video using a touch-sensitive display. In this latter fashion, for example, virtual "breadcrumb" waypoints for navigation could be laid out in regions lacking sufficient scene contrast for natural landmark following, such as outdoors in desert terrain. Alternatively, the operator could illuminate attributes of interest with a laser pointer, such as the targeting laser on *ROBART III's* weapon, for example. One of the first mobile systems to actually do this was *Hermies IIB*, developed at Oak Ridge National Laboratories, where remote operators would designate small objects for the robot to pick up with its manipulator, using a tripod-mounted laser synchronized to the video capture system on the robot.[10] In this example, the vision-centric frame of reference is used to not only direct the motion of the robotic platform in approaching the object, but also to control the actions of the robot's manipulator in picking it up.

In similar fashion, such a *vision-centric* scheme lends itself nicely to high-level weapon control. *ROBART III* maintains a pre-taught database of digital color pictures of potential targets. The vision system compares these target templates with live images from its incoming video stream, using a color-correlation-matching algorithm which operates on the Red, Green, and Blue color channels of each image. Intensity values are normalized so that algorithm performance is independent of brightness level, and correlation results which exceed a specified threshold are considered matches. (The current algorithm requires that the target distance and perspective be similar to that of the template images, but size- and angle-independent methods are under investigation.) Once strong correlation is detected, an approximate vector to the target is computed, and both the pan-tilt-zoom camera and weapon can be generally trained accordingly.

For each target-type, an image database of vulnerable locations associated with that target is also stored. The camera is zoomed in to obtain a high-resolution image, whereupon the same correlation-based matching is performed, and results with high correlation are considered areas of vulnerability. For example, Figure 7 presents four sequential zoomed-in images of a cardboard box situated on the seat of an office chair. The box itself serves as the pre-taught target, while within the box one or more soda cans represents pre-taught vulnerabilities for where best to shoot this particular target.

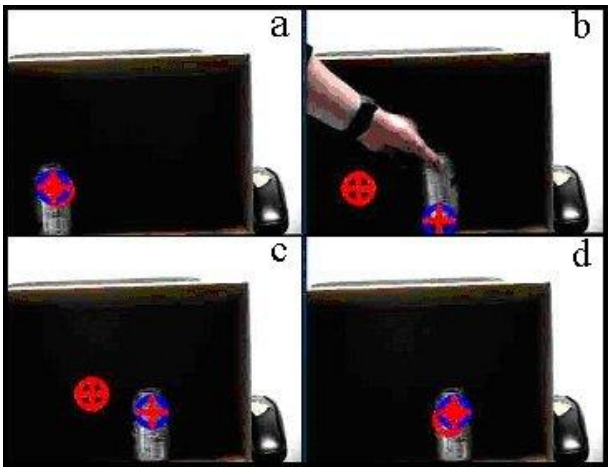

Fig. 7. a) Targeting laser on detected vulnerability (soda can); b) Can is tracked in real-time while being relocated; c) Targeting laser servos to new location; d) Laser now relocated on new target position, ready to fire weapon.

Training *ROBART III's* Gatling-style gun on these perceived vulnerabilities implies a high degree of precision, possibly imaging just a few square centimeters at distances of tens of meters, which normally would require precise calibration between the respective weapon and camera coordinate systems. We present a solution which is extremely accurate but requires no special calibration, wherein a bore-sited laser is used to assist in the final target acquisition. The laser is cycled on and off in sync with the frame rate of the vision system, so that the portion of the scene illuminated by the laser only shows up in every other capture frame. Simple image subtraction reveals the precise location of illumination, whereupon the error vector between the laser spot and the desired point of impact is calculated. A closed-loop control

algorithm servos the weapon until the laser footprint coincides with the perceived vulnerability, and firing accuracy is virtually guaranteed.

5.3 Model-Centric Reference

An even higher-level frame of reference would be some type of absolute world model to which both the human and the robot could relate. The most obvious example here is GPS coordinates, since both humans and robots use the GPS system for outdoor navigation. SSC San Diego's MOCU controller, for example, can input ortho-rectified overhead imagery (or electronic nautical charts in the case of unmanned surface craft) to serve as the map display, and automatically converts user-selected visual waypoints on the display into GPS-coordinate waypoints for use by the robot.[5] The operator thus can draw a detailed route using his mouse or stylus on a situational display, or simply indicate the desired goal destination and let the robot figure out how to get there.

An alternate approach for waypoint designation in the field, developed by Exponent, Inc., San Diego, CA, employs a laser rangefinder with an integrated directional compass (Figure 8). The operator sights through the viewfinder of the hand-held instrument, centers the crosshairs on a particular point of interest, then presses a button to capture measured range and bearing. This information is then passed to the OCU via a short-range Bluetooth RF link, where it is combined with current GPS position and used to calculate the GPS coordinates of the observed location. Exponent's approach is attractive in the sense that it can be implemented without adding much additional weight, if any, as the warfighter typically will already have binoculars, a compass, and in some cases a rangefinder, which would no longer be needed as separate items.

Fig 8. Exponent's prototype waypoint-designation system.

In indoor applications, however, GPS is not available, and so the robot must reference using range data from its surroundings. The MDARS-Interior robot navigated in this fashion using a system of virtual paths developed by John Holland of Cybermotion, with enhancements for off-path transit added by Gilbreath.[11] While arguably appropriate for fixed-site security installations, such reliance upon a priori maps is not practical in tactical applications, where the requirement exists to enter structures of opportunity with no advance knowledge of the interior layout. Simultaneous localization and mapping (SLAM), a relatively new approach to indoor localization undertaken within the research community over the past several years, now enables robotic platforms equipped with laser and/or stereo ranging systems to build an accurate map as they explore an unknown environment, and to keep themselves localized within that map at the same time.

For example, SRI International (SRI) has developed (under the DARPA TMR program) a very efficient technique called *Consistent Pose Estimation* that incorporates new laser range information into a growing map representation, successfully addressing the problem of *loop closure* (how to optimally register laser data when the robot circuitously returns to a previously mapped area).[12] Their approach builds upon and further enhances algorithms using a representation of the robot's state space based on Monte Carlo sampling.[13] Accordingly, in FY-04, SSC San Diego tasked the Idaho National Engineering and Environmental Laboratory (INEEL) under a Memorandum of Agreement to assist in porting over and enhancing the SRI SLAM solution, incorporating in the process INEEL's own robust collision avoidance scheme.[14] FY-05 efforts will also address the addition of a global path planner to operate upon the SLAM-generated world model.

By arbitrarily growing the laser-generated 2-D SLAM map some finite vertical amount as shown in Figure 9, we can essentially create a 2½-D virtual reality representation that in many cases provides more situational awareness for driving than actual real-time video imagery (in part because today's younger generation feels very much at home with such video-game perspectives).[15] This is important not only for improved effectiveness, but also because the required RF bandwidth is significantly reduced if video is no longer required.

From the perspective of natural-language control, the issue now becomes one of semantics (i.e., how does one exert voice control over SLAM-based motion?). The path planner expects any requested destination to be identified in terms of its X,Y location in the map, normally provided by an operator simply clicking on the desired goal location in the map

display. Verbally instructing the robot to "go to coordinate 1203, 1856" is not only awkward and error prone, it is also non-intuitive. The operator would have to place the mouse cursor over that location in the display to first determine the destination coordinates, at which point a simple click could then provide the coordinates the old fashioned way.

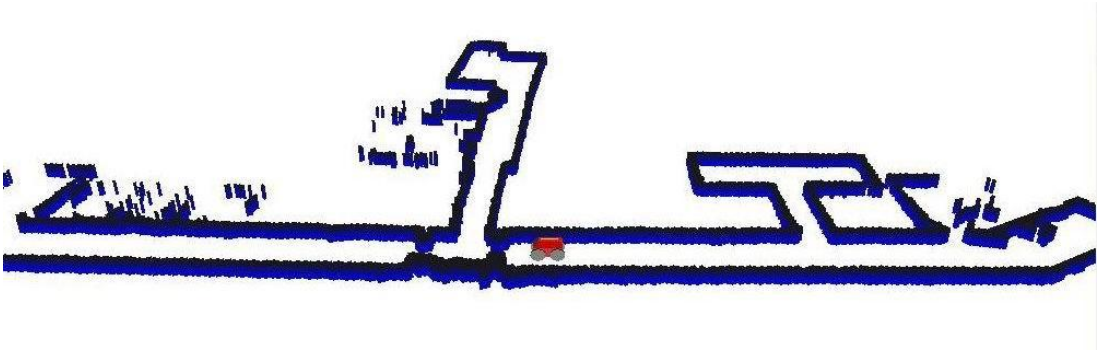

Fig 9. SLAM map of Battery Woodward, an underground WW-II bunker at SSC San Diego, after the robot has made one pass down each of the main hallways, and further explored one T-shaped room on the right.

Furthermore, recalling the earlier example where the robot received an e-mail instructing it to go to the sender's office and take a picture, how does commanded motion, once parsed, get conveyed to the path planner? The simplest way would be to have the parser glean the destination office from the e-mail sender's identity, and use a pre-supplied look-up table to convert the doorway to that particular office into an X,Y coordinate for the path planner. But this approach relies on a priori information, and our goal for the *Warfighter's Associate* is to assume there is none.

In addressing this problem (as will be specifically illustrated later), we have elected to exploit the concept of *augmented reality*. But first, examining the taxonomy proposed by Milgram,[16] it can be seen in Figure 10 that the *Reality-Virtuality Continuum* is defined on the one end by the *real environment* (as would be reflected in the robot's video), and on the other by the *virtual environment* (as is represented in the SLAM model). Conventional work in augmented reality typically links to the video image additional information describing the scene under observation, to provide the viewer more complete situational awareness. For example, if the video includes a view of a certain building, any available amplifying information about that building automatically appears in the image, most simplistically in the form of a pop-up text overlay. Such amplifying information could come from a variety of different sources, to include other warfighters in the area, deployed robots, overhead imagery, or other conventional intelligence sources. In such fashion, for example, a warfighter observing the video could instantly be alerted to the fact that this particular building had been swept and found free of snipers, contained radiological contamination, and had three subterranean levels. For this reason, *augmented reality* displays are seen as having tremendous potential for increased situational awareness on the battlefield, and this technology is therefore considered key to the *Warfighter's Associate* concept.

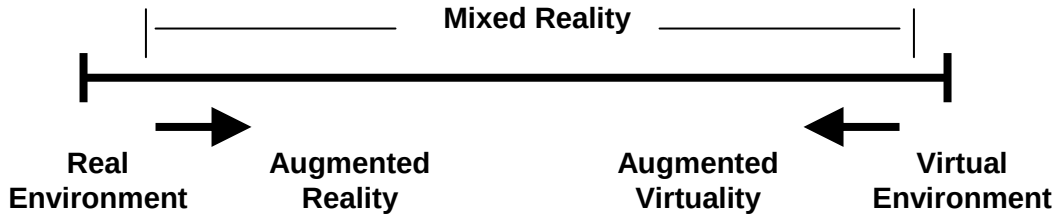

Fig. 10. The *Reality-Virtuality Continuum* as proposed by Milgram.

It was previously mentioned, however, that one key advantage of our incredibly robust SLAM navigation is the reduced (if not eliminated) need for real-time video. Accordingly, instead of pursuing the conventional *augmented reality* approach (which a number of very competent research organizations are already addressing), we intend in the near term to work primarily on the other end of Milgram's scale, in the realm of *augmented virtuality*. That is to say, rather than augment the video with additional detail, we will instead augment the virtual world model (derived from SLAM) with even more virtual information (derived from on-board sensors and/or human input). The former (conventional)

approach requires very precise registration (of robot position, orientation, and camera gaze) with an expanded virtual model representing the additional information to be overlaid, which is problematic. With the latter approach, additional data collected by the robot's sensors can simply be time- and position-stamped with respect to the SLAM model, making registration (at least of robot-collected data) rather simplistic indeed.

For example, we have already developed chemical, gas, and radiological sensor payloads that have in fact been used in Iraq, albeit upon purely teleoperated robots. If these same sensors were mounted on an autonomous robot that could explore an underground bunker complex using SLAM, any associated sensor readings over pre-set alarm thresholds could be "tagged" with an appropriate icon in the *augmented-virtuality* layer of the SLAM world model. If desired, video and/or still imagery (see Figure 11) of conditions encountered at that location could similarly be linked to this same tag, for later viewing upon demand. The robot conducts a comprehensive autonomous building sweep, augmenting its virtual world model with any relevant data uncovered in the process, and requiring no direct human supervision at any point in the entire process.

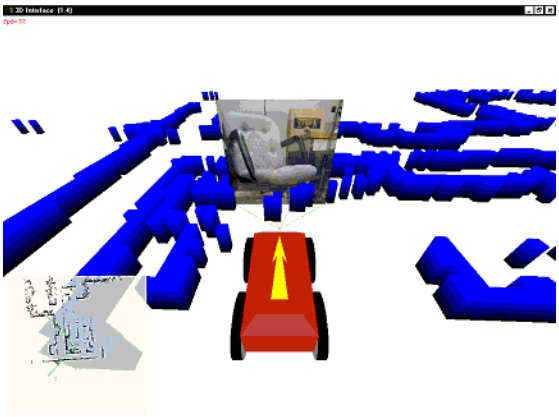

Fig 11. Screenshot of the virtual model fused with real-time video images from an ATRV robot exploring INEEL office space.

Note that in this scenario, the robot's video camera is treated as just another onboard sensor that can contribute snippets of reality to the virtual model, versus the other way around. If the vision system detects a doorway, it could apply a heuristic to examine the wall to either side to see if there was an associated room identification sign, as shown in Figure 12. Close examination of the sign yields a room number (and in this case the principle occupant), which can then be added to the augmented virtual model. Exploiting this simplistic vision primitive allows the robot to "learn" a tremendous amount of information about its previously unknown environment. Furthermore, these room identification "tags" can be specifically associated with the X,Y coordinates of the doorway center, thus enabling verbal direction of the robot to a specific SLAM model location, in the form of: "Go to Room 102," without the need for a priori lookup tables or direct specification of location coordinates. If room identification signs are not conveniently posted, identification labels can be supplied by the operator (Figure 13), or arbitrary room numbers sequentially assigned by the robot as the rooms are discovered.

Fig 12. *ROBART III's* vision system finds the room sign and associated identification number.

Fig 13. Information overlays augmenting the virtual model, such as text labels for room identification, can also be directly entered by the operator.

6. AUTONOMOUS BEHAVIORS

It is important to recognize that implementing a natural-language interface that allows the operator to provide high-level direction implies the robot must also be equipped with the appropriate autonomous functionality to execute the requested actions. More simply put, you can't provide high-level direction to a system that does not know how to execute high-level behavior. This section describes some of the intelligent behaviors already implemented, and identifies some future candidates under follow-on consideration.

6.1 What do we have working now?

Basic mobility commands – Low-level drive commands, such as move forward, turn left, slow down, stop, etc.
Guarded mobility commands – *ROBART II* introduced reflexive teleoperation in 1990, using screen icons to support high-level oversight for driving with low-level sensor assist. The initial implementation included doorway seek/penetration behaviors (i.e., find and enter doorways to left, right, or in front), with automated camera pan-axis coordination (so the view matched the commanded action), facilitating much less demanding operator supervision.[9]
Explore and map structure – Simultaneous localization and mapping (SLAM). While *ROBART III* had a rudimentary capability for simultaneous localization and mapping,[17] vastly superior schemes were developed by SRI[12] and Carnegie Mellon University (CMU)[13] under DARPA's TMR program.[4]
Building sweep – An extension of "explore and map" that incorporates looking for specific items of interest, such as human presence, chemical agents, radioactive contamination, etc.
Surveillance – Providing remote video back to the human operator.
Static motion detection – Detection of object motion from a stationary robot. This is a well developed behavior that originated on *ROBART I* in the early eighties, was significantly enhanced on *ROBART II* using a combination of video, optical, acoustic, vibration, passive infrared, and microwave sensors, then ported over to the MDARS-Interior robot for production installation. More recently, full 360-degree peripheral video coverage, originally developed as part of the Distributed Interactive Video Array (DIVA) project,[1] was incorporated on *ROBART III.*
Static motion tracking – The ability to track a moving target from a stationary platform. Automatic camera panning was initially demonstrated on *ROBART II*, transferred to the MDARS-Interior robot, and then further refined on *ROBART III*. Full two-axis pan-tilt-zoom tracking developed under DIVA was incorporated in January 2004.
Weapon control – The ability to use the target-tracking solution to aim and fire a weapon. *ROBART III* expanded the reflexive-teleoperation concept employed on *ROBART II* to include automated weapon control, with limited video-based target tracking (pan axis only). Full-frame two-axis tracking developed under DIVA was adapted to support interactive target acquisition and weapon firing in February 2004.
Vehicle or people following – The ability to generate platform-motion commands from the target-tracking solution in order to follow a moving person or vehicle. *ROBART II* used a sonar-based approach to follow a human at walking speeds, as long as there were no intervening obstacles. A video-based tracking system (embedded in the Sony *EVI-D30* pan-tilt-zoom camera) is used to support "people following" on our iRobot *All Terrain Robotic Vehicle (ATRV),* with the concurrent ability to avoid intervening obstacles that may appear between the *ATRV* and the target being followed.[4]
Convoy capability – An extension of "vehicle following" that allows an intelligent robot to lead a convoy of much less sophisticated robots or vehicles for a variety of missions. An initial near-infrared beacon-tracking algorithm was demonstrated in 1998, using LynxMotion *Hexapods* slaved to *ROBART III* (Figure 14). SSC San Diego's DARPA-funded Autonomous Mobile Communication Relay project later employed laser tracking of retro-reflective targets and added an RF-repeater functionality to ensure extended communications when operating in complex structures.[18]
Automated recharging or refueling – The ability to automatically connect to a recharging (or refueling) system to replenish the onboard energy source. Both *ROBART I* and *ROBART II* could activate an optical beacon via an RF link for automatically locating their recharging system. The MDARS-Interior robot used a near-infrared beacon (and later sonar) for connecting with its charger. The UGV/UAV integration effort, which calls for a three-phase development approach for the automated launch, recovery, refuel, and relaunch of a UAV from a UGV, is now in its second phase. Vision-based docking to a free-standing recharging stationed is planned for *ROBART III.*

Fig 14. Four LynxMotion *Hexapods* in trail behind *ROBART III*.

Send/receive e-mail – The ability to receive simple text commands via an e-mail connection, and to transmit text information or requests, with appended attachments (mpeg video, jpg still imagery, etc.) as needed.
Simplistic sign interpretation – The ability to identify room numbers posted near doorways in an indoor environment, and subsequently generate an associated text label which corresponds to the doorway X,Y location in the SLAM model.

6.2 What are we going to do next?

GPS waypoint navigation – The ability to proceed from current GPS position to a new GPS location or waypoint. As *ROBART III* was initially intended for indoor operation only, the intent is to port over the existing GPS waypoint navigation scheme from the *URBOT* to allow indoor/outdoor operation.[5]

Visual landmark homing – The ability to visually track a reference landmark while the vehicle is moving, for purposes of maintaining a constant heading. JPL demonstrated this capability under the TMR program, and SSC San Diego is implementing a similar functionality with concurrent obstacle avoidance.

Motion detection on the move – Detection of moving targets while the robot itself is in motion. There are at least four ongoing efforts in the research community that SSC San Diego is monitoring: 1) particle filter with vision (Perceptek); 2) particle filter with ladar (University of Washington);[19] 3) passive radar (University of Texas Applied Research Lab), and 4) simultaneous location and mapping with detection and tracking of moving objects (CMU).[20]

Complex sign interpretation – The ability to identify signs along the roadside or in/on buildings, zoom in, and perform character recognition to generate a text string. The result can then be passed to the text parser and analyzed for information potentially useful for increased situational awareness (i.e., town names, highway numbers, distances to next town, building numbers, floor numbers, exit locations).

Respond to hand gestures – The ability to visually perceive and interpret common hand gestures for traffic control, such as an upraised hand for stop, a waving forward gesture to proceed, pointing to the left or right to indicate a desired change in direction. Thus if the robot encounters a traffic cop at an intersection or accident scene, then that person could visually influence the robot's path without having any RF connection for voice or control data.

6.3 What's further down the road?

Face recognition – The ability to zoom in on a detected human presence, capture a few frames of the facial area, and compare to a database of pre-taught images to identify as friend, foe, or unknown. Many research efforts are already pursuing this issue for homeland security and other related applications.

License plate capture – The vehicular equivalent of "face recognition" above, also being widely addressed.

Foreign language interpreter – The ability to recognize foreign speech, translate, and then speak the English equivalent (or vice versa). The Army's Applied Research Lab (ARL), Adelphi, MD, is already pursuing an ability to scan in foreign language documents, translate, and then print out in English in the field.

Cognitive awareness – The ability to infer additional information from context to influence the interpretation of a verbal command, or to serve in the absence of a command altogether. A police dog, for example, can sense when its handler is injured, and take steps to protect or go for help.

7. CONCLUSION

This paper has introduced the concept of a *Warfighter's Associate*, with an intentional focus on historical development of the various supporting technologies, in an attempt to show that near-term feasibility in some bounded capacity is more practical than one might otherwise have thought. Two underlying technology thrust areas are involved: 1) a natural language interface that allows the robot to receive high-level verbal commands, and, 2) the ability of the instructed robot to then execute high-level autonomous behaviors. In the latter case, we rely heavily on the OSD-funded *Technology Transfer Program* to improve the autonomy and functionality of candidate platforms.[4]

The initial emphasis of our effort has focused on high-level exploration of interior structures, where SRI's SLAM and INEEL's collision-avoidance algorithms have effectively solved the navigational issues for single-floor scenarios. Remaining technical challenges to fully autonomous indoor navigation include the ability to climb stairs, as well as the ability to open doorways, since unfettered access to all interior spaces is not very representative of real-world conditions. As both of these issues were explored by various researchers under TMR, we have elected not to directly address them, but concentrate instead on an *augmented-virtuality* expansion of the SLAM model. We will eventually extend this concept and our associated focus to exterior applications, evolving a *Warfighter's Associate* prototype that is not limited in scope to one environment or the other, but able to freely transition back and forth as required.

8. REFERENCES

1. Everett H.R., "Robotic Security Systems," *Instrumentation and Measurement Magazine,* IEEE, Vol. 6, No. 4, pp. 30-34, December, 2003.
2. Everett H.R., "Autonomous Navigation on a Shoestring Budget," *The Robotics Practitioner*, (http://www.spawar.navy.mil/robots/land/robart/shoestring.html), pp. 15-23, Winter, 1996.
3. Everett H.R., Gilbreath G.A., Tran T.T., Nieusma J.M., "Modeling the Environment of a Mobile Security Robot," NOSC Technical Document 1835, Naval Ocean Systems Center, June, 1990.
4. Pacis Estrellina, Everett H.R., "Enhancing Functionality and Autonomy in Man-Portable Robots," Proceedings, SPIE Unmanned Ground Vehicle Technology VI, Defense and Security, Orlando, FL, 12-16 April, 2004.
5. http://www.spawar.navy.mil/robots/land/r3v/r3v.html
6. Avizonis Pepi, "Advanced Robotic Controller: Phase 5 Program Report," Exponent, Inc., SD10012.PH5.0104.0147, January, 2004.
7. Zimmer Carl, "Mind over Machine," *Popular Science,* pp. 46-52, 102, February, 2004.
8. Weizenbaum Joseph, *Computer Power and Human Reason*, New York: Freeman, 1976.
9. Laird R.T., Everett H.R., "Reflexive Teleoperated Control," Association For Unmanned Vehicle Systems, 17th Annual Technical Symposium and Exhibition (AUVS '90), Dayton, OH, pp. 280-292, July-August, 1990.
10. Kilough S.M., Hamel W.R., "Sensor Capabilities for the HERMIES Experimental Robot," ANS Third Topical Meeting on Robotics and Remote Systems, Charleston, SC, CONF-890304, Section 4-1, pp. 1-7, March, 1989.
11. Holland J.M., Martin A., Smurlo R.P., Everett H.R., "MDARS Interior Platform," Association of Unmanned Vehicle Systems, 22nd Annual Technical Symposium and Exhibition (AUVS '95), Washington, DC, July, 1995.
12. Gutmann J.S., Konolige K., "Incremental Mapping of Large Cyclic Environments," in *Proc. IEEE Intl. Symp. On Computational Intelligence in Robotics and Autonation (CIRA)*, 1999.
13. Fox D., Burgard W., Dellaert F., Thrun S., "Monte Carlo Localization: Efficient Position Estimation for Mobile Robots," Sixteenth National Conference on Artificial Intelligence, (AAAI'99), July, 1999.
14. http://www.inel.gov/adaptiverobotics/autonomousbehaviors/motion.shtml
15. C.W. Nielsen, B. Ricks, M.A. Goodrich, D. Bruemmer, D. Few, and M. Walton. "Snapshots for Semantic Maps." Proceedings of 2004 IEEE Conference on Systems, Man, and Cybernetics. October 10-13, 2004, The Hague, The Netherlands.
16. Milgram P, Kishino F, "A Taxonomy of Mixed Reality Visual Displays," IEICE Transactions on Information Systems, Special Issue on Networked Reality, Vol. E77-D, No. 12, December, 1994.
17. Everett H.R., Gilbreath G.A., Ciccimaro D.A., "An Advanced Telereflexive Tactical Response Robot," *Autonomous Robots*, Vol. 11, No. 1, Kluwer Academic Publishers, pp. 39-47, July, 2001.
18. Nguyen Hoa, Pezeshkian Narek, Gupta Anoop, Farrington Nathan, "Maintaining Communication Link for a Robot Operating in a Hazardous Environment," ANS 10th International Conference on Robotics and Remote Systems for Hazardous Environments, Gainesville, FL, March 28-31, 2004.
19. http://www.cs.washington.edu/ai/Mobile_Robotics/mcl/animations/floor3D.avi
20. Wang Chieh-Chih, Thorpe Chuck, Thrun Sebastian, "Online Simultaneous Localization and Mapping with Detection and Tracking of Moving Objects: Theory and Results from a Ground Vehicle in Crowded Urban Areas," Proceedings, IEEE International Conference on Robotics and Automation, May, 2003.