Urban Short Range Interaction: An LVC Solution for Urban Operation Training

Tijmen Muller, Robbert Krijnen, Gillian Visschedijk
TNO
Kampweg 5, 3769 DE Soesterberg, The Netherlands
{ tijmen.muller, robbert.krijnen, gillian.visschedijk }@tno.nl

ABSTRACT

Urban Operations are an increasingly important part of military operations, both nationally and during expeditions. The complexity of the urban environment makes these operations difficult, and a key aspect of this complexity is the presence of local population. The individual soldier needs to constantly analyze unclear situations, predict civilians' intentions and be able to rapidly make decisions in order to both ensure their own safety and prevent lethal mistakes, which would endanger the goodwill of the population. Gaining experience with civilians through training is essential for successful execution of urban operations.

In the Urban Training programs of the Royal Netherlands Army live training facilities are available, but a varied group of well-trained role players is scarcely available and costly to use. This paper presents the result of a three year research project into an innovative enhancement to live training for Urban Short Range Interaction (USRI). In this concept, trainees enter a live, physical environment (a room) enriched with virtual role players (projected on the wall) that respond directly to the user's actions.

A technical demonstrator was developed that integrates commercial-off-the-shelf elements, such as a gesture and a speech recognition system. We report on the user value analysis that was carried out with various parties in the Defense and Safety domain, presenting the phases in the training programs where the application of USRI has the most training value. Finally, we describe the system requirements that have been evaluated with the intended users.

ABOUT THE AUTHORS

Tijmen Muller is scientific researcher at the Training and Performance Innovations department at TNO. He graduated at the University of Twente in Enschede with a degree in computer science. He is a simulation and serious gaming expert and has specialized in the design and architecture of simulation, research into advanced simulation techniques and the use of simulation and serious gaming for concept development, experimentation and training, in particular for the military and safety domains.

Robbert Krijnen is member of the scientific staff in the Defence, Security and Safety Division at TNO since 1996. As a lead engineer and technical consultant he has been involved in a wide range of projects within the military operations domain. He is specialized in the field of enabling technologies in the Modeling, Simulation and Gaming domain. Robbert holds a Masters of Science Degree in Computer Science from the University of Delft in the Netherlands, with a specialization in Computer Graphics.

Gillian Visschedijk is scientific researcher at the Training and Performance Innovations department at TNO. She graduated cum laude at the University of Twente in Enschede with a degree in educational science. Gillian specializes in the use of serious gaming and simulation technology for training purposes, mostly in the military and safety domain. Her focus is on the didactical measures needed for application.

Urban Short Range Interaction: An LVC Solution for Urban Operation Training

Tijmen Muller, Robbert Krijnen, Gillian Visschedijk
TNO
Kampweg 5, 3769 DE Soesterberg, The Netherlands
{ tijmen.muller, robbert.krijnen, gillian.visschedijk }@tno.nl

INTRODUCTION

Urban operations are an increasingly important part of military operations because of the urbanization world wide and the increasing number of irregular conflicts (Hart, 2008). Already in 1999, NATO predicted that the majority of expeditions are expected to take place in urban environments (NATO-RTO, 1999). Also, urban areas are the main operational environment for public safety organizations such as the civil and military police. The urban environment not only poses physical challenges, such as complex terrains and obscure dangers, but also the challenge of dealing with the presence of local population.

In current military operations, especially in peace-keeping and counter-insurgency (COIN) operations, the population may well be the key to success. Winning the hearts and minds makes preventing collateral damage a key factor, which is complicated by the fact that opposing forces are not always recognizable as such. These opposing forces tend to blend into the civilian population, are familiar with the local environment and culture and are willing to use unusual tactics, such as suicide attacks. This makes irregular forces the biggest threat in urban environments.

Consequently, there is a great responsibility on the individual level. On one side, every soldier has to try to guarantee the safety of their own forces, while on the other side, he has to be reticent in the use of force in order to prevent civil casualties which may have strategic or even political consequences – this responsibility is reflected by such terms as *junior leader* and *strategic corporal* (Krulak, 1999). The same holds for individuals responsible for public safety, such as police officers or Special Weapons and Tactics (SWAT) teams. Experience shows that an unnecessary casualty during an arrest may be followed by social unrest and may also lead to political consequences.

These conflicting goals (own safety and preventing collateral damage) require situation assessment and decision making under time pressure. It is essential to give trainees the best possible training by immersing them in these challenging situations. Nowadays, several training facilities for urban operations are available. For example, in the Netherlands there are two military training sites, 'Oostdorp' and 'Marnehuizen' (see Figure 1) and the police training site 'Ossendrecht'. However, most of the exercises on these sites are aimed at the procedural level. Training for situation assessment and decision making skills need populated urban training sites. Sometimes human role players are used, but this has several disadvantages:

- Role players are not always available.
- Role players are costly: hiring professional role players is expensive, while using trainees as role players takes away training value.
- The quality of role playing is not always guaranteed: if role play is not performed by well-instructed professionals, it may not be delivered in a way that brings the highest training value. Additionally, training quality may be reduced by human limitations such as fatigue, illness or just having a bad day.
- Available role players are limited and homogeneous: especially when trainees are used for role play, they are typically young, white males. The use of some types of role players, such as elderly and young children, may be hard to provide, ethically unacceptable or even restricted completely because of legislation.
- Using non-military role players may be restricted, because of the confidentiality of trained procedures (e.g. Special Forces or Air Marshalls).
- The use of human role players may be impossible

Figure 1. Live training facility Marnehuizen

in some situations, for instance when using live ammunition.

Advancing developments in the field of sensors and simulation provide a solution to these restrictions: using *virtual role players*. Within the Dutch research program "Integration of Live, Virtual and Constructive Simulation" a demonstrator was developed with this purpose: *Urban Short-Range Interaction* (USRI).

THE CONCEPT

The goal of USRI is to present the user with a virtual role player with whom he can interact in a meaningful and non-intrusive way in a live training environment. This means that the trainee is aware of the virtual role player, but also the other way around: the virtual role player needs to know what the trainee is doing in order to react appropriately. Our research focuses on the value of such a training system for various domains and the drawbacks and necessary developments of such a live, virtual and constructive (LVC) solution. First, we developed a technical demonstrator, with requirements based on interviews held with a number of stakeholders (most importantly the Instruction Group for Urban Operations and Simulation Centre Land Forces of the Royal Netherlands Army). Next, the demonstrator was evaluated by stakeholders in terms of suitability for training goals and additional requirements.

Our research focuses on: 1) keeping the user interaction as natural and the system as non-intrusive as possible: reducing the individual-worn equipment and allowing the user to operate his own equipment; 2) proposing an extended set of observable actions and reactions for a responsive computer-controlled virtual role player in a variety of scenarios; and 3) using commercial and government off-the-shelf technologies and components, creating a potentially low-cost and reliable system.

Presenting virtual role players to human users requires integrating a live environment with a virtual environment. Three classes of LVC training systems can be distinguished, each with its advantages and disadvantages:

1. individual-worn virtual reality: head-mounted display, so the user is immersed in a purely virtual world;
2. individual-worn augmented reality: a see-through head-mounted display, where the virtual world augments the real world;
3. facility-based mixed reality: a virtual world projected in a real room.

An inherent limitation of the first, purely virtual, system is that the user loses a large part of the interaction with the physical world. Since the observed virtual world is intangible, the user cannot lean against walls or open doors. Also, users need to be translated into the virtual world very precisely to allow physical communication like signaling by touching a team mate's shoulder. Similarly, all intended actions (i.e. moving, weapon use, hand signals) by a user need to be translated to the virtual world, otherwise this information is lost. This mapping of the real world onto the virtual world is either done by placing sensors on the user or by providing interaction devices (e.g. a joystick) for these actions. An advantage of a purely virtual system is that it is relatively easy to vary the presented simulated environments.

A technical limitation that holds for both individual-worn systems is the limited field of view of current head-mounted displays. As a result, the user cannot sense any movements in the periphery of his visible environment (JTAA, 2010). The robustness and cost of the currently available head-mounted systems are also a point of concern.

A general challenge for all three classes is that the computer-controlled virtual role player needs to be responsive to a level that makes them believable. Otherwise realism, immersion and, arguably, training effectiveness is violated (JTAA, 2010, 2011). This implies that the virtual role player has to be conscious of the user and his actions (e.g. the user aims his gun at the virtual role player) and is able to react to these actions. Firing a weapon is one of these actions but given the context of urban and peace-keeping operations, less kinetic and non-lethal interactions are also needed, such as verbal and non-verbal communication.

Given our focal points mentioned above, the USRI system belongs to the third class, where the real world is not replaced but enhanced by a virtual world. USRI extends a real room by presenting one or more virtual role players, for example as projection on one or more walls of the room. We chose to use a virtual role player instead of a physical presentation of a role player, as a virtual role player can be made interactive with relative ease. A physical presentation of a role player is either static, such as a pop-up target or a dummy, or very hard to animate, as with animatronics. Consequently, a virtual representation of a role player can more easily give varying and realistic reactions.

A variety of sensors placed in the room or on the equipment gathers information on the actions of the user(s), such as his position, weapon use and speech. The user may freely move in this instrumented room, not limiting the physical aspects of the procedures for entering a room. The gathered information is processed by the virtual role player and used to select one of many possible actions, resulting in an interaction between user and virtual role player (see Figure 2).

TRAINING GOALS

The USRI concept is beneficial for several training programs. From military training for urban operations to police training for close protection or SWAT teams. Although the usage of USRI will definitely differ between these training programs (as explained below), at its essence all of them are about 'action intelligence'. Within the military or police domain, action intelligence is a term to explain the difficulty many trainees face when they are presented with a real situation with real people and real danger and still need to have the legal framework or rules of engagement in mind. Accordingly, it is one thing to know what to do in dangerous situations, being in the middle of one with adrenaline rushing through your body is something very different. Some trainees simply freeze while others may make decisions based on inadequate situation assessments, for example, due to tunnel vision. Tunnel vision in the context of situational awareness is the situation in which someone only concentrates on one thing (e.g. a guy holding a knife) while paying too little or no attention to everything else (e.g. he was actually cooking). This can even mean that the trainee does not follow procedures and focuses on his sector, disregarding other potential threats in the environment. Both freezing and tunnel vision are well-known psychological phenomenon when operating under stress and as both can have major consequences, it is of vital importance to prepare trainees sufficiently. Two training applications with different training goals are described in more detail to illustrate the USRI concept.

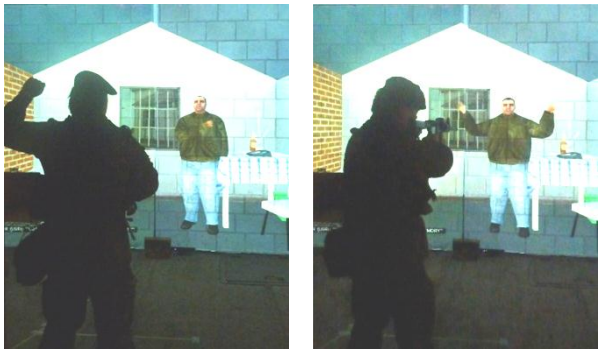

Figure 2. The virtual role player reacts differently on the user's actions

The first example application is to use USRI in a so called 'shooting house'. At one point in the urban operations training program, military trainees are placed in a situation where live ammunition is used instead of blanks or FX non-lethal ammunition. As a trainee enters a house together with his team, this is already a highly serious and exciting experience. Human role players are, as you can imagine, out of the question here. It is very important for the trainees to cope with situations where, for example, they have to clear a house while they know in advance that it may become dangerous. Once inside the house, they meet a hostile person in one room and deadly force is needed to control the situation. The team then continues to the next room where they meet an anxious mother and her child who have no hostile intentions. It is difficult for the trainee to keep his head cool and not attack the mother and child while in a very aroused state of mind expecting more hostile people in the house. In other words, making appropriate situation assessments and decisions to either shoot or not shoot in only a split second repeatedly is very demanding. USRI instantiates these situations by representing different types of virtual entities in different rooms, with various behaviors and intentions, while at the same time trainees are aroused by the use of real bullets when entering the house with their teammates.

Instead of the 'shoot-no shoot' decision described above, more nuanced decisions have to be made by the civil or military police. Subsidiary and proportional behavior is required, which means that a) the least aggressive weapon suitable for the situation should be used; and b) the way of acting is appropriate within the legal framework (i.e. Rules of Engagement). In other words, the training goal is to learn what is appropriate *and* legally allowed, and this under highly stressful situations. The most straightforward example of such a situation is when a trainee needs to arrest an aggressive person. The trainee constantly needs to mentally keep a birds eye view, assess the situation and possibly switch between different levels of violence, for example when the person pulls out a knife after a short quarrel. Some trainees think they can still solve the situation verbally, while another trainee would already have shot the suspect. Allowing civil and military police trainees to experience these types of situations and to reflect on them afterwards with the legal framework as a background is extremely valuable. To make this happen, the USRI concept needs to be extended with different types of simulated weapons, gestures and enriched voice recognition. Its potential value has

however already been recognized by both civil and military police training schools.

The two example training experiences illustrate the possible training goals which can be met by using USRI: split-second decision making, shoot-no shoot decision, subsidiary and proportional use of violence and situation assessment with bird eyes views. Not only do the training schools support these new applications of the USRI concept, but they even proposed the training goals themselves. Further possibilities for implementation and the adjustments to the concept are currently being discussed.

DESIGN

Designing the USRI demonstrator presented two major challenges. The first challenge is representing different types of humans in a realistic way at different locations within a house. The second challenge deals with the interaction between the users and virtual humans.

For the first challenge, representing the virtual humans, we looked at different visualization techniques and work already done in similar applications. One of the major design goals was to minimize the instrumentation of the users: we wanted them to wear as little equipment as possible. The main reason was to prevent the users from knowing something was going to happen because they had to change equipment or start using a Head Mounted Display (HMD) and thus keeping the tension of the mission as high as possible. During the design stage we were not sure what kind of scenarios were going to be used. However, we anticipated that the user should be able to get close to the virtual human and should be able to perceive the physical and emotional state of the virtual human, e.g. see a nervous twitch or observe an obscured hand.

The second challenge deals with the interaction between the user and virtual player. The application needs to know what the user is doing in the environment so the virtual player can react. In order of importance we composed the list below. The top six requirements were classified as ‘must haves’, the remaining requirements are ‘nice to have’.

1. Voice recognition. Being able to respond to simple commands (e.g. show your hands, lay down).
2. Equipment handling (e.g. firing, state and direction of weapons).
3. Non-intrusive (marker less) tracking techniques to minimize user instrumentation.
4. Tracking 2 to 4 users at the same time in one room.
5. Knowing the position and orientation of the user.
6. User identification. Who is who and who is in which room.
7. Gunfire feedback. Communicating results of hostile (role player) gunfire on the different users.
8. Detecting user postures (skeleton with position of legs, arms and hands), e.g. standing, kneeling.
9. Detecting user’s gestures (skeleton movement), e.g. greeting, waving.
10. Accurate hit point detection. Know which body part the user fired at.
11. Tracking the direction of view (head). This is necessary to recognize who the user is talking to when presented with multiple virtual humans.
12. User characteristics (holding weapon, wearing sunglasses, helmet). How the user is perceived by the virtual human can make a big difference e.g. wearing a helmet is more aggressive than a beret.
13. Eye tracking. Together with the direction of view this information is important to detect if the user covered his sector e.g. should have been able to detect possible threats.
14. Facial expressions of the users.

Because the application is targeted for use in existing urban training facilities we also looked at incorporating the existing instrumentation capabilities. The Dutch army uses the Mobile Combat Training Center (MCTC) from Saab. This laser-based simulation system provides the capability for tactical live training in urban areas, enabling direct and indirect fire. The GPS equipped system is mainly meant for outdoor use and has limited capabilities for indoor tracking. If rooms are instrumented it can only report which user is in which room. Because the detail of position tracking is too low, the system was discarded as a possible solution. However for user identification and gunfire feedback (red on blue) the MCTC system can be used, as discussed in the next section.

IMPLEMENTATION

The main design goal to minimize the instrumentation of the users while providing a natural way of interacting with the virtual role player presented many challenges. This section presents the lessons learned while implementing the visualization (simulation platform and goal presentation) and the interaction (speech recognition, weapon use and user tracking).

Simulation platform

The implementation of the USRI application was targeted on Virtual Battlespace 2[1], with several reasons to do so. VBS2 is generally accepted as one of the best COTS simulations for dismounted soldier training currently on the market and contains a large amount of useful content, such as a variety of entities (male, female, children, animals) and objects to be used in the virtual environment (e.g. suspicious objects such as weapons and IED components). Additionally, it is relatively easy to make VBS2 interoperable with external components using the advanced scripting interface (ASI). A final, pragmatic reason is that the Royal Netherlands Army has an enterprise license for VBS2, including a terrain database of the Marnehuizen training facility, making it a logical choice for research.

The decision to use VBS2 meant that the visualization and interaction with the users was not centered on a normal desktop computer game-play with a monitor, keyboard and mouse. To allow this difference in game-play a client-server setup was used. The server handles all the external interactions, while a VBS2 plug-in functions as a client to the server to receive commands and translate these commands into VBS2 script actions. The use of VBS2Fusion was considered as it promises an advanced programming interface. The features of interest within the context of this project were the direct application programmable interface (API) and 'skeletal control'. Unfortunately, the advertised skeletal control was non-existent. The API allowed us to directly interact with VBS2 without using the indirect scripting interface. This increased the performance considerably but decreased the flexibility in the development of the plug-ins. This is acceptable for the development of a well-defined interface but not in a prototyping setting for this project. Concluding, the advantages of VBS2Fusion where not sufficient compared to our existing client-server architecture already in-place.

A disadvantage of the VBS2 system is the lack of low-level control for the virtual role player. For example, the number of possible facial expressions is very limited (only six moving facial points: 3 for the mouth and 3 for the eyebrows); the entity's body cannot be animated in real-time, but only by using pre-recorded animations; and the behavior of the unit can be unpredictable – for example, when the entity is told to lay down on the floor, it would sometimes start crawling away from the user, creating an unintended event. Depending on the training goal and the level of the trainee, when these nuances become important in the decision making process, the aforementioned issues may trigger an improper interpretation (for example: "He is not showing any emotion, so he must have been expecting us and is not afraid.") or may make the virtual role player unbelievable.

Designing a scenario for USRI basically came down to deciding how each virtual role player should react given the actions of the users. A reaction to a user entering the room can for example be changing its stance, trying to flee, pulling a gun or threatening the user verbally. As more users' actions are recognized and the virtual role player has more (re)actions available, the number of possible courses within a scenario grows rapidly. In this case, it may advisable to develop a behavior model for each virtual role player, as scripting each scenario course becomes laborious.

Presentation of the virtual role player

Considering the goal of reducing the instrumentation on the user, the number of options for presenting the virtual role player is limited. Because a head-mounted display is no option and the virtual role players need to be presented in a believable way, we choose to project them on a maximized display area. A typical display area from the floor up to at least 2.5 meters was needed to display a human with hands-up appropriately. Depending on the situation, smaller display surfaces can be used. For example, simulating a window or physically limiting the required display size by using real furniture e.g. a desk or cabinet in front of the virtual human, as in the FlatWorld system (Pair, 2003).

Because the user should be able to get close to the virtual human and we wanted to demonstrate the application in existing training facilities we selected a front projection solution with projectors capable to being positioned close to the 'projection' wall (i.e. short-throw projector). A vertical mounted high-definition projector was used to project an image of approximate 1.5 meters wide and 2.8 meters high. This is sufficient to display a full-size virtual human. Two projectors where used to create an area of 3 by 2.8 meters for increased immersion. The high brightness (2500 lumens) of the selected projector was sufficient to display the scene on a non-prepared wall (e.g. the gray bricks of a room in one of the training facilities). If reducing tunnel vision is a learning goal, it should be possible to present multiple virtual role players in a single room in order to confront the users with multiple threats. This is possible by using several projectors, but reduces the accessible area – using a rear-projection solution could solve this problem, as described below.

[1] http://www.vbs2.com

We considered using a stereo projection solution using passive filter eyeglasses like used in many 3D cinemas. A stereo projection solution enables us to let the user determine relative depths in a perceived scene. These depth cues are important to the human brain to determine if objects (weapons) are hidden behind objects or a body at close range. The illusion of depth is created by presenting two offset images separately to the left and right eye of the viewer. The different filters for the left and right eye in the eyeglasses filter the offset images to each eye. Using circularly polarized glasses instead of linear polarized glasses enables the viewer to tilt his or her head and still maintain left/right separation. These glasses are relatively low cost. The disadvantage is that the user must wear the glasses to be able to perceive the depth effect. When not wearing the glasses, the image will be perceived with distortion due to seeing the offset images for the left and right eye simultaneously. When wearing the eyeglasses in a normal setting (daylight), the eyeglasses function as light shaded sunglasses. However, these glasses offer no protection to sunlight (UV A/B) and will hinder the user when entering a low-light environment making them want to take them off. Last but not least, the glasses offer no protection from the use of blanks close by, which is mandatory in live urban training.

To keep the setup simple and relatively low-cost, we decided that the added value was too limited and sense of depth could be created using other techniques like viewer position dependent images (off-axis projection) and larger display areas. A floor to ceiling and wall to wall setting like in a CAVE does give a greater sense of immersion than stereo projection.

When creating new training facilities with virtual human interaction in mind, it is advisable to use a rear-projection solution. This maximizes the area where the users are able to go without blocking the projection. Also environmental conditions can be better controlled within a small enclosed room with only the projection/computer equipment. The additional costs are relatively low when integrating a rear-projection solution within a new building. Technology improvements with auto-stereoscopic techniques will also enable the use of glasses-free 3D in the future.

With the developments in the area of holographic projections, future work could include the presentation of a virtual role player in the middle of a room. This increases realism, as trainees are now able to move around the virtual role player, and makes training more complex, as trainees have to consider their position in the room even more: when able to encircle the virtual role player, the risk of blue-on-blue is increased.

Speech recognition

The aim of using automated speech recognition (ASR) is to allow the user to give simple commands to the virtual role player, such as “Turn around”, “Get down” and “Show your hands” or some variation. These commands should be able to be recognized within a typically noisy environment using a headset for each user. Different voice recognition solutions where tested to see which performed best.

The Automated Language Training System (ALTS) from Alelo[2] was considered briefly, but this product is targeted at intercultural competency and foreign language skills. The functionality of real-time dialog with conversational agents was of a higher level than required. The ASR module used by Alelo is based on the open source engine Julius[3]. The functionality of Julius seemed sufficient for this application and was tested instead of the ALTS. The disadvantage of the system is that for each command ‘command templates’ have to be trained, preferably by many different users to cover all possible pronounciations for the specific commands. Only variations in pronunciations included in the training can be recognized. So if the user pronounces “Show me your hands” it will not match the trained template “Show your hands”.

A second system that has been evaluated is Loquendo[4]. This system is used mainly in interactive services systems like telephony-based contact automations. Compared to the Julius solution, the system uses pre-trained telephone speech acoustic models. The utterances that have to be recognized can be defined in a grammar, which specifies all possible ways a command can be pronounced. Given an uttered command, the system returns a number of recognized options with a probability for each. This way, the system supports multiple options in uttering a certain command, without the need to train the system beforehand. It also supports multiple environment profiles (telephony, mobile, automotive) to maximize performance in relation to environment noise. We tested with the “telephony” profile and expect to improve performance by using other profiles. However this was not tested within this project. A disadvantage of this system is that it is aimed at native English speakers: when used by Dutch speaking English, the recognition probability is reduced somewhat.

There are several factors that influence the success rate of automated speech recognition:

[2] http://www.alelo.com/
[3] http://julius.sourceforge.jp/en_index.php
[4] http://www.loquendo.com/

1. Complexity of the grammar: the smaller and less complex the commands the ASR needs to recognize, the better the recognition will be. Within the USRI system the commands that have to be recognized are limited. However, when speech is recorded which was not meant to be recognized (for example, when users are talking to each other), the ASR will still try to recognize this, returning the 'best' solutions, each with small probability. Hence it is important to configure the ASR in a way that these solutions will not pass the probability threshold and will be discarded; otherwise, another way needs to be found to distinguish between talking to the virtual role player and all other speech.
2. Number of speakers: the complexity of the recognition task increases when the system cannot be optimized for a small, fixed set of speakers. This is the case for USRI, as the system is expected to be used by large numbers of trainees, making speaker-dependent recognition impossible.
3. Background noise: background noise, especially other voices, will decrease ASR performance. This problem can be reduced by recording voice as close to the speaker as possible, for example by using a headset, as we have done for USRI. However, this is less desirable, given the goal of designing a non-intrusive system.
4. Spontanity of speech: less spontaneous speech increases the success rate, as the articulation is usually more precise. Within USRI the user typically has an informal and excited voice, increasing difficulty of recognition. However, at the same time the user may try to make himself explicitly clear in communication towards the role player, which may improve recognition.
5. Information on start and end of a command: if the system exactly knows which part of audio actually is relevant, the success rate will increase. A possibility is to manually tag the audio. For USRI we could use a so-called press-to-talk solution, but this decreases the natural interaction which is one of the initial goals of the project. Therefore automatic recognition of the beginning and end of a command is used, adding to the difficulty of recognition and also creating a small delay.

Instead of using personal headsets as we have now, a microphone could be set up in a fixed place in the room, making the system less intrusive. We expect the recognition to worsen, but it also becomes unclear which of the users is actually speaking. Speaker recognition can now be done by identifying the used microphone, but with a fixed microphone we could use the microphones integrated in the Kinect camera, which should be able to detect which direction the sound is coming from. This information on who is speaking can be used by the virtual role player to know which user is talking to it, which could affect its behavior.

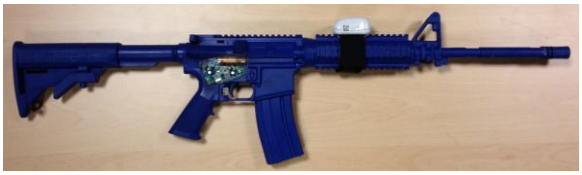

Figure 3. Instrumented mockup of a Diemaco C7

The technology used for speech recognition is relatively mature and does not seem to promise a giant leap in improvements within the next couple of years. There are still some new commercial developments (e.g. 'Siri' from Apple), but they have to improve considerably to be successful. Additionally, it is uncertain whether this technology will be available as a stand-alone system. Nevertheless enhancements within these systems may trigger new improvements in competitive systems.

Weapon system

For the instrumentation of the equipment used by the users, we used a mockup of a Diemaco C7 assault rifle (see Figure 3). No real version of the weapon was used due to legal limitations. The weapon was instrumented by integrating a trigger mechanism connected to a wireless Bluetooth mouse. The internal electronics of the mouse was completely integrated in a space within the weapon. With this capability, we were able to detect trigger events and respond accordingly.

Only knowing the trigger event is not sufficient to determine the fire effect, so an orientation tracker (Intersense BT[5]) was mounted on top of the weapon. With an accuracy of ±½ degrees pitch and ±1 degrees yaw it should be able to do coarse hit point detection. Knowing the orientation of the weapon also allowed us to let the virtual humans react to the aiming of the weapon. Thus entering a room with a raised weapon can have a different effect on the virtual human then entering the room with a lowered or holstered weapon.

Even knowing the orientation of the weapon is not enough to be able to do coarse hit point detection. The position of the weapon also plays an important role. When knowing the position of the user (see section on user tracking below) there have to be made assumptions on the location of the weapon compared to the user's position. For this application we made the

[5] http://www.intersense.com/pages/18/60/

assumption that the end of the rifle will be roughly at the same position in front of the user, thus independent of left handed or right handed users. At close range, this assumption works reasonably well. An auto-aiming functionality was added to VBS2 for usability purposes; however for accurate hit point detection, the exact position of the weapon has to be known. Instead of also measuring this position, future work could include a permanent activated laser (laser sight) for detecting the exact aiming direction.

Future improvements involve instrumenting a real weapon with a minimal number of sensors: an integrated fire-activated laser (activation based on muzzle flash and/or sound) and a permanent laser for detecting the aiming direction. Instrumenting a real weapon enables the use of air, non-lethal and live ammunition to increase the realism of the training and allows the trainees to operate their own weapon. It is desirable to make use of Saab's MCTC system, which is already used in field training exercises, to make integration of USRI in these live exercises possible.

User tracking

Considering the goal of reducing the instrumentation on the user, the number of options for user tracking is limited. We considered the following options:

- Inertial based tracking
- Ultra-wideband tracking
- Depth sensing camera tracking techniques
- Marker based tracking

Inertial based tracking An inertial based tracking system uses motion sensors (accelerometers) and rotation sensors (gyroscopes) to continuously calculate the position, orientation, and velocity (direction and speed of movement) of a moving object without the need for external references. A major disadvantage of these tracking systems is that they typically suffer from accumulated error. Because the system is continually adding detected changes to its previously-calculated positions, any errors in measurement, however small, are accumulated from point to point. This leads to 'drift', or an ever-increasing difference between where the system thinks it is located, and the actual location. Therefore the position must be periodically corrected by input from some other type of navigation system.

For tracking human movements (walking, running) the brief periods of time the velocity and acceleration of the foot are zero can be used to determine the drift error and be used to correct the acceleration data, thus limiting the overall error. When the human makes a more sliding movement this principle will not work.

For temporary position and orientation tracking in GPS-denied areas, this tracking technology is a good solution with a minimal instrumentation of the user. However for detecting postures multiple tracking devices would be needed, increasing the instrumentation and posture detection algorithms.

Depth sensing camera tracking With the introduction of the Kinect motion sensing device by Microsoft, for the Xbox 360 game console, there have been many developments with camera based tracking techniques. Using a depth-sensing camera is a low-cost solution to our goal of designing a non-intrusive system, as the camera can collect information on user's actions without the need to equip the user with sensors. We implemented the Kinect camera in the USRI system, using the OpenNI SDK and PrimeSense's NITE middleware (as this was the only SDK available at the time).[6] The NITE middleware identifies users and tracks their movements, and provides the framework API for implementing Natural-Interaction user interface controls based on gestures. The OpenNI provides a device-independent API for writing the application.

The OpenNI/NITE framework allows tracking of up to four persons. The sensor data allows tracking of the position, orientation, stance and gestures. This information can be used by the virtual role player to respond, for example, to the presence of a user, the number of users in the room, the orientation of a user (do something when he is not paying attention) or specific gestures, such as greeting or pointing.

Considering the moment of entering a room, the Kinect in combination with OpenNI does not need calibration, as a user profile can be saved offline and reloaded on each entry. However, it is not guaranteed that a user leaving and reentering the room will be recognized as the same person. If this information is needed for the behavior of the virtual role player, additional markers on the user will probably be needed.

The use of a tool, such as a weapon, cannot be independently tracked, but is recognized as part of the user, making independent tracking of tools necessary, as described before for the weapon system. Additionally, the recognized position of the user (and therefore, the tool) is not detailed enough to do precise hit point detection: it is not possible to detect where a user has fired based on just the user position and the weapon orientation.

[6] http://www.primesense.com/nite

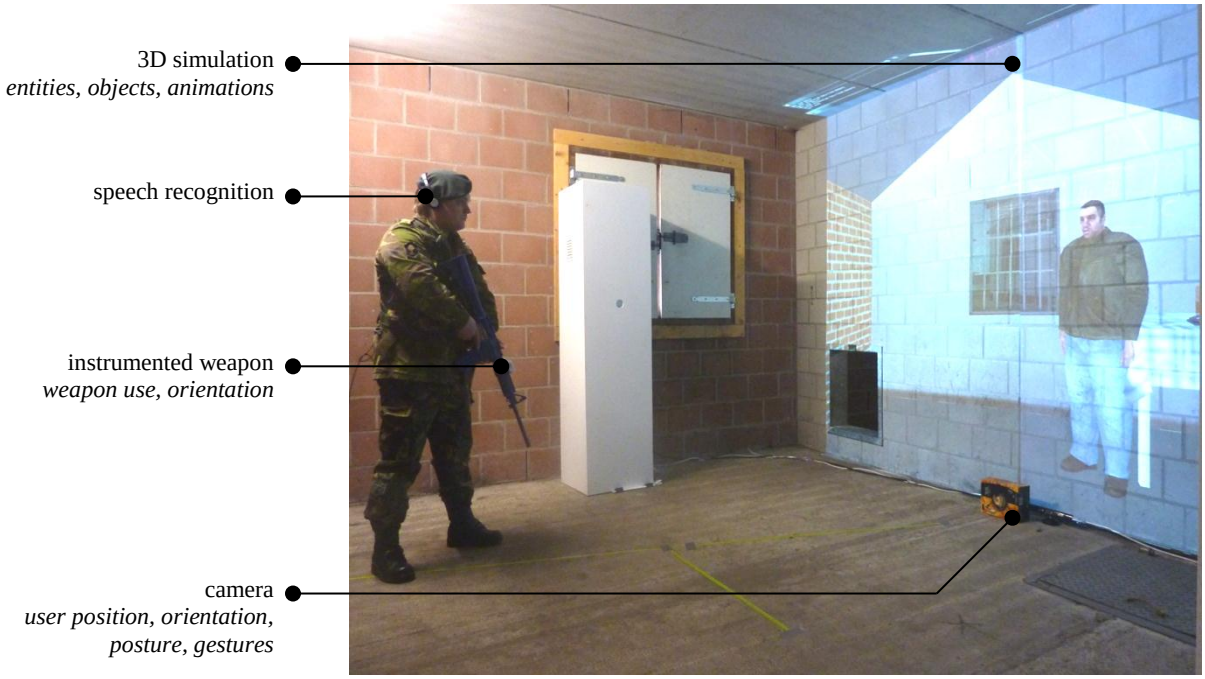

Figure 4. USRI demonstrator in the live training facility Oostdorp

The camera can only track users in a limited area (between 0.8m and 4 meters with a 57° horizontal field of view and 43° vertical field of view). This may be sufficient for user tracking in a small room, but this becomes problematic in larger rooms or non-square rooms. A possible solution is to link multiple cameras, increasing the area in which users can be tracked. The currently available SDKs (OpenNI and Microsoft) do not support the simultaneous use of multiple Kinect devices. When using multiple Kinects, controlled by multiple computers, there are device dependent limitations to consider. This has to do with the way the depth sensing works. An IR structured light (pattern of light points) is projected by the Kinect and an IR camera captures this pattern. Based on the location and size of the light points a depth image is produced. Consequently the pattern of light projected by one Kinect may not overlap the pattern of another Kinect, because the software will be unable to function correctly when patterns from other sources are also captured. Summarizing, multiple Kinects can be used as long as they do not cover the same area. This limits the use of the Kinect. Other camera based tracking systems should be used to cover larger or odd sized rooms. Interesting results can be found in current academic research (Luo, 2010): a video-based 3D human motion recognition based system uses new techniques for 3D tracking of people and understanding their poses even when the users occlude each other.

Since depth-sensing cameras have only recently been picked up commercially by large corporations like Microsoft, we expect quick improvements in the near future. The fact that Microsoft now provides an easy-to-use SDK[7] free of cost boosts the third-party development of applications outside of the initial Xbox scope and these applications are usually made available through an open-source community.

Marker based tracking The marker based tracking solutions use active or passive markers in combination with cameras to determine position and possibly the orientation of users. Although this is an intrusive tracking technique it was considered for user identification. Placing a small specific marker on a user can uniquely identify a user. However, the position of the marker and cameras play an important role in the accuracy and was considered too complex to implement in existing facilities.

A special case of markers is the use of RFID tags. These markers can be read from a distance using a special reader. The distance from which a tag can be read is determined by the reader's antenna size and required power output. This solution can be used to equip all users with a unique RFID and place readers near entry points in a room. With multiple readers the position of RFID markers can also be triangulated very accurately. These systems (Ultra-Wideband tracking systems) are well-suited to short-distance applications and indoor use. However, like with the cameras, the required setup of multiple radios was considered too

[7] http://www.microsoft.com/en-us/kinectforwindows/

complex to implement in existing facilities. Future work has to determine if the use of RFIDs with readers near the entry point could be used to identify a user and link it to the camera based motion tracking system.

EVALUATION

The system as depicted in Figure 4 was evaluated by a variety of stakeholders: civil and military police (including close protection teams), the Instruction Group for Urban Operations (IGUO), marines and special forces. As can be expected from the varying backgrounds and training goals, requirements differ for the mentioned units.

The IGUO believes the system can already be of value in shooting houses and possibly in their training facilities to train shoot/no-shoot decision making. Their minimum requirements are: usable by up to 4 persons; virtual role player reacts to shots fired and voice commands; and projected virtual role player should be well visible in daylight conditions.

The usefulness within the training facilities depends on the mobility of the system, as it cannot be set up permanently in these facilities: setup time should not exceed 2 hours. A typical exercise is on company level and is aimed for two 'suspect' houses. Two rooms per house need to be instrumented, so a minimum of 4 USRI systems is needed to support an exercise in the training facilities. Within shooting houses, mobility is not an issue, since the projectors and sensors may be part of a permanent setup there.

For training the use of subsidiary and proportional force, there are additional requirements. On the one side the user should be able to choose how to act in each situation by using body language, speech, non-lethal weapons or lethal weapons, including direction of fire. This means that specific stances need to be recognized (e.g. signaling the virtual role player to stop or to lay down), that non-lethal weapons need to be modeled and that direction of fire (aiming for legs, chest or head) needs to be registered in detail. Additionally, more subtle behavior is needed from the virtual role player in order to present a higher variety of scenarios. This requires more animations and the simulation of facial emotions.

The use of USRI for training close protection (personal security) tasks is less clear, since the focus is typically not on entering a room, but usually involves identifying suspicious behavior within large crowds. However, two applications of interest have been identified. The first is training on exfiltration: when escorting a VIP within a building, one might end up entering a room and running into an unexpected individual, which could be presented by the USRI system. A second application is the use of USRI within the building in which the close protection teams usually operate. As they often operate in specific buildings, our approach can be used to repeatedly train in this environment.

Finally, reducing tunnel vision as a learning goal requires the presentation of virtual role players on multiple locations in the room. This may be an initially obscured location when entering the room, allowing the virtual role player to move away from the user.

REFERENCES

Hart, 't, M., Vink, N. & Buiel, E. (2008). *An Introduction to Urban Operations,* TNO-D&V A338, Den Haag: TNO.

Joint Technology Assessment Activity (JTAA) (2010). *Future Immersive Training Environment (FITE) Joint Capability Technology Demonstration (JCTD), Operational Demonstration 1, Independent Assessment Report.*

Joint Technology Assessment Activity (JTAA) (2011). *Future Immersive Training Environment (FITE) Joint Capability Technology Demonstration (JCTD), Operational Demonstration 2, Independent Assessment Report.*

North Atlantic Treaty Organization, Research and Technology Organization (NATO-RTO) (1999). *Land Operations in the Year 2020, RTO Technical Report 8.*

Krulak, C.C. (1999). *The Strategic Corporal: Leadership in the Three Block War.* Marines Magazine. Retrieved June 11, 2012, from http://www.au.af.mil/au/awc/awcgate/usmc/strategic_corporal.htm

Luo, X., Berendsen, B., Tan, R.T. & Veltkamp, R.C. (2010). *Human Pose Estimation for Multiple Persons Based on Volume Reconstruction.* 20th International Conference on Pattern Recognition (ICPR).

Pair, J., Neumann, U., Piepol, D. & Swartout, B. (2003). *FlatWorld: Combining Hollywood Set-Design Techniques with* VR. IEEE Computer Graphics and Applications