An Instrumentation and Computational Framework of Automated Behavior Analysis and Performance Evaluation for Infantry Training

Hui Cheng, Rakesh Kumar, Chumki Basu, Feng Han, Saad Khan, Harpreet Sawhney, Chris Broaddus, Chris Meng, Azhar Sufi and Thomas Germano
Sarnoff Corporation
Princeton, NJ, USA
{hcheng, rkumar, cbasu, fhan, skhan, hsawhney, cbroaddus, cmeng, asufi, tgermano}@sarnoff.com

Mathias Kolsch and Juan Wachs
Naval Postgraduate School,
Monterey, CA, USA
{kolsch, jpwachs}@nps.edu

ABSTRACT

Infantry training, ranging from basic training to joint exercises can become more effective through automated performance evaluations and performance based real-time training adaptation and control. In this paper, we will introduce an automated behavior analysis and performance evaluation computational framework that we developed based on United States Marine Corps (USMC) Tactics, Techniques and Procedures (TTP) for a wide range of training objectives. We will also discuss the experimental results of instrumented training systems using this framework for data collection, performance evaluation and multimedia after-action-review (AAR).

We model warfighters' behavior (individually and in teams) as *states,* and the causes of state transition as *trigger-events*. Each state has a set of performance metrics. Both states and trigger events have detectors. TTP are represented as hierarchical Finite State Machines (FSM) with associated performance metrics. Training exercises are constructed as a set of preplanned trigger events to exercise all states defined in the training objectives. Behavior analyses use sensor data observations to estimate states, and performance evaluations compute performance metrics given the estimated states of the trainees. We also develop a novel Histograms of Oriented Occurrence (HO2) algorithm for individual and team action recognition.

We instrumented training systems for both outdoor urban operations and indoor close-quarter battle. Video cameras capture the training exercises and automatically analyze behavior and evaluate performance. Each warfighter's location, weapon pose, and head orientation is tracked using a combination of video-based people tracking, GPS, RFID (Radio Frequency ID), video analysis and inertia navigation sensors. Gunshots are captured through trigger sensors. Our system estimates behaviors and corresponding performance metrics in real-time, and ingests those data into a database. Events and videos are overlaid on a 3D-model of the training site to enhance AAR and situational awareness, and furthermore, AAR allows searching and browsing of training events and the computation of statistics.

ABOUT THE AUTHORS

Dr. Hui Cheng is the Technical Manager of the Adaptive and Cognitive Systems Group at Sarnoff Corporation. He received his Ph.D. degree in Electrical Engineering from Purdue University, West Lafayette. Dr. Cheng's research interests are in the areas of computer vision, video based behanviro analysis, pattern recognition, artificial intelligence, statistical image modeling and simulation and training. Dr. Cheng has published more than 32 articles and holds 8 U.S. patents. He is a senior member of IEEE, the Chair of Princeton/Central Jersey Chapter, IEEE Signal Processing Society and a member of IEEE Technical Committee on Multimedia Systems and Application.

Dr. Rakesh "Teddy" Kumar is the Senior Technical Director of the Vision and Robotics Laboratory at Sarnoff Corporation. Prior to joining Sarnoff, he was employed at IBM. He received his Ph.D. in Computer Science from the University of Massachusetts at Amherst in 1992. His technical interests are in the areas of computer vision, computer graphics, image processing and multimedia. He was an Associate Editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence from 1999 to 2003. He has co-authored one book on Video Registration, more than 50 research publications and has received over 22 patents.

Dr. Chumki Basu is a Member of Technical Staff at Sarnoff Corporation. She is a Program Manager for projects in the areas of simulation and training, Internet-based computer vision applications, and artificial intelligence. She is also a technical lead in the areas of knowledge representation and language analysis. In the past, she was a Research Scientist at Bell Communications Research (Applied Research division). Her interests include ontologies, multi-media information retrieval, text mining, machine learning, recommender systems, and user modeling. Dr. Basu holds a Ph.D. degree in Computer Science from Rutgers University

Dr. Feng Han is a Member of Technical Staff in the Adaptive and Cognitive Systems Group at Sarnoff Corporation, Princeton, New Jersey, USA. His research interests are in the areas of computer vision, pattern recognition, machine learning. He has published over 17 articles. He is a member of IEEE. Dr. Han holds a Ph.D. degree in Computer Science from the University of California, Los Angeles.

Dr. Saad M. Khan is a Member of Technical Staff in the Adaptive and Cognitive Systems Group at Sarnoff Corporation, Princeton, New Jersey, USA. His research interests are in the areas of computer vision, pattern recognition and statistical image modeling where he has published over 15 articles. He is a member of IEEE, and serves as the Chair of IEEE Princeton Central Jersey Section Signal Processing Society. Dr. Khan holds a Ph.D. degree in Computer Science from the University of Central Florida

Dr. Harpreet S. Sawhney is a Technical Director, leads the Vision & Learning Technologies Lab. at the Sarnoff Corporation. Dr. Sawhney received his Ph.D. in CS at UMass, Amherst, in Computer Vision. His areas of interest are Video Indexing, Motion Video Analysis, Object and Event Recognition, 3D Modeling, Vision & Graphics Synthesis, Data Mining and Compact Video Representations. Dr. Sawhney has been an Associate Editor for the IEEE T-PAMI, and has also served on the Program Committees of numerous Computer Vision and Pattern Recognition conferences. He has published over 75 papers and holds 22 patents.

Mr. Chris Meng is a Member Technical Staff of the Software Group at Sarnoff Corporation.. Prior to joining Sarnoff, he was with Video Insight, Inc. He received his M.S. in Computer Science from the University of Houston Clear Lake. His technical interests are in the areas of digital video surveillance application, SQL database application, GUI, C# and VB.NET.

Mr. Tom Germano is a Member of Technical Staff in the Vision and Learning Laboratory at Sarnoff Corporation. Prior to joining Sarnoff, he was at Vizta3d developing demonstration computer graphics applications. He received his Bachelor of Arts in Computer Science from New York University in 2001. His technical interests are in the areas of real-time computer graphics, computer vision, and multimedia.

Mr. Azhar Sufi is a Member of Technical Staff in the Embedded Vision Group department at Sarnoff Corporation. He graduated with a B.Sc. degree in Electrical Engineering from Virginia Tech in 2005 and has been with Sarnoff since. He is currently pursuing his M.Sc. degree at Rutgers University in Electrical and Computer Engineering.

Dr. Mathias Kölsch is an Assistant Professor of Computer Science at the Naval Postgraduate School. He is also affiliated with the MOVES Institute, Chair of the MOVES Academic Committee and the Academic Associate for the MOVES curriculum. His research interests include computer vision, hand gesture recognition, augmented and virtual environments, sensor networks and mobile/embedded computing.

Dr. Juan Wachs is an Assistant Professor at the Industrial Engineering Department at Purdue University. Dr. Wachs was a postdoctoral fellow at the Naval Postgraduate School at the Computer Science Department where he worked on problems related to body posture recognition, sub-resolution tracking and surveillance applications. He completed his PhD on Intelligent Systems in Hand Gesture Vocabularies Design, concentrating mainly on pattern recognition, cognitive and physiological measures for hand gesture interfaces for robot control.

An Instrumentation and Computational Framework of Automated Behavior Analysis and Performance Evaluation for Infantry Training

Hui Cheng, Rakesh Kumar, Chumki Basu, Feng Han, Saad Khan,
Harpreet Sawhney, Chris Broaddus, Chris Meng,
Azhar Sufi and Thomas Germano
Sarnoff Corporation
Princeton, NJ, USA
{hcheng, rkumar, cbasu, fhan, skhan, hsawhney, cbroaddus, cmeng, asufi, tgermano}@sarnoff.com

Mathias Kolsch and Juan Wachs
Naval Postgraduate School,
Monterey, CA, USA
{kolsch, jpwachs}@nps.edu

INTRODUCTION

The success of ground operations, especially in an urban environment against an asymmetric threat, heavily depends on the effectiveness of infantry training. In the last decade, military training has been improved in many aspects, from range modernization to the integration of live, virtual and constructive (LVC) training technologies. Many training facilities today are covered by video cameras and radio frequency (RF) localization system for capturing and tracking trainees' locations and for improved performance evaluation and after-action-review (AAR). Although a large amount of video and tracking data are collected, the status quo for performance evaluation is manual grading by instructors. This lack of automated behavioral analysis and performance evaluation has also prevented closed-loop training that adapts to training requirements in real-time and provide more realistic and challenging interactions at the level of individual needs.

Infantry training, from basic training at home stations to joint exercises prior to deployment, can become more effective through automated behavior analysis and performance evaluations. In this paper, we will introduce an automated behavior analysis and performance evaluation computational framework for a wide range of training objectives. This framework was developed for military operation in urban terrain (MOUT) based on US Marine Corps Tactics, Techniques and Procedures (TTP).

In the proposed behavior analysis and performance evaluation framework, we model warfighters' behavior (individually and in teams) as *states*, and the causes of state transitions as *trigger-events*. Each state has a set of performance metrics. We developed an Augmented Hidden Markov Model (AHMM) for estimating both states and trigger events. A TTP is captured as hierarchical Finite State Machines (FSM) with associated performance metrics. A training exercise or a training scenario is constructed as a set of trigger events to exercise all states defined in its training objectives. We also developed a training ontology for MOUT and represent both TTP and exercises using RDF (Resource Description Framework).

Our behavior analysis module uses sensor data as observations to estimate the states that the trainees are in. The performance evaluation module computes the performance metrics given the estimated states of the trainees. We also develop a novel Histograms of Oriented Occurrence (HO2) algorithm for individual

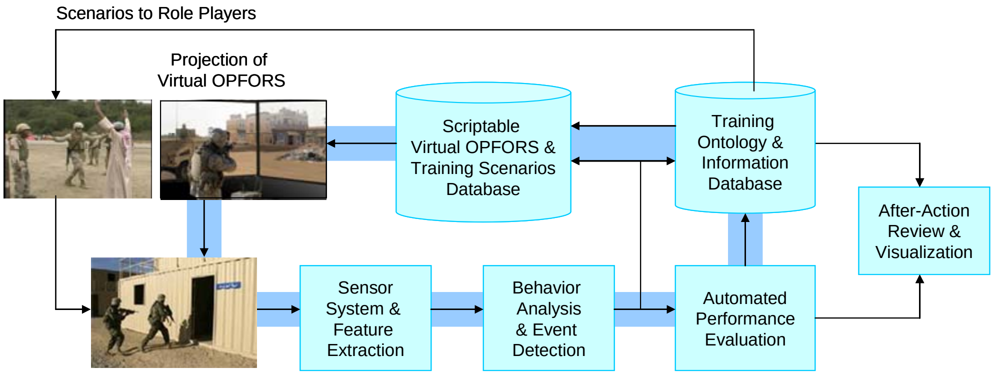

Scenarios to Role Players
Projection of Virtual OPFORS
Scriptable Virtual OPFORS & Training Scenarios Database
Training Ontology & Information Database
After-Action Review & Visualization
Sensor System & Feature Extraction
Behavior Analysis & Event Detection
Automated Performance Evaluation

Figure 1. System block diagram.

and group activity recognition that captures the interactions of multiple players in one feature vector.
We instrumented training systems for both outdoor urban operations and indoor close-quarter battles (CQB). The system can use either role players or projected synthetic virtual Opposing Forces (OPFORS) or both as stimuli. Video cameras capture the training exercises and automatically follow trainees for behavior analysis and performance evaluation. Each warfighter's location, weapon and head orientations are computed using a combination of GPS or RFID, inertia navigation sensors (INS) data and video analysis. Gunshots are captured through trigger sensors. Videos and detected events are overlaid on a 3D-model of the training site for enhanced AAR and situational awareness experiences. Additionally, our AAR allows searching and browsing of training events and the computation of statistics. Our system estimates behaviors and corresponding performance metrics in real-time, and ingests both into a database. This prototype training system has been implemented and applied to mock Marine Corps MOUT exercises, and experimental results and subjective participants' feedback have shown improved training efficiency and effectiveness as a result of the system.

SYSTEM ARCHITECTURE

The prototype training system that we developed may be applied in both indoor and outdoor MOUT training. It can also use both virtual OPFORS projected on walls as well as role players as stimuli for an exercise.

As shown in Figure 1, the prototype training system has six major components: (1) *Sensor System and Feature Extraction* module to capture and compute each warfighter's location, head and weapon pose, actions (e.g., a trigger pull), the location of shots, and the impact of the shots on participants (including virtual characters); (2) *Behavior Analysis and Event Detection* module to estimate the state that the participants are in, and to detect and recognize events of interest based on the sensor system's outputs; (3) *Training Ontology and Information Database* to capture expert knowledge (including TTP of MOUT operation), and storage of each warfighter's location, pose, action, detected events and performance metrics for AAR; (4) *Automated Performance Evaluation* to compute performance metrics for each individual and the entire team, and to detect any mistakes made during an exercise in order to index them in the database for fast retrieval and advanced AAR capabilities; (5) *After-Action-Review and Visualization* to provide both an iconic view of movement, actions and events in a 3D environment and a synchronized video display by combining all video feeds onto a 3D-model of the MOUT environment; and finally, (6) *Scriptable Virtual OPFORS & Training Scenarios Database* for use when virtual OPFORS are used for an exercise (generally for indoor exercises), so that one may control the behavior of projected OPFORS by concatenating behaviors in this database.

AUTOMATED BEHAVIOR ANALYSIS AND EVENT DETECTION

Automated behavior analysis and event detection is a key component for an advanced training system. Detection and recognition of arbitrary human behavior is an extremely challenging problem because of the vast number of possibilities in an unconstrained environment. However, for training applications, behavior analysis is greatly simplified due to the controlled environment, the staged stimuli and the set of expected behaviors already known to the system. Taking advantage of the training domain, our behavior analysis framework (Figure 1) uses a finite state machine (FSM) model where participants' behavior are the *states* and the transitions of states are caused by stimuli that we refer to as *trigger events*. The goal of behavior analysis is to estimate the states of the participants and the states that the participants should be in at any given time. The former are used for exercise and scenario control and the later are used for performance evaluation. To robustly detect each state, we build classifiers for not only for each state, but also for each trigger event. At a given time, based on the state estimation, a set of related classifiers are activated for detecting trigger events and states that can be transitioned to and from the current states.

Behavior Analysis Framework

We model a training exercise as a finite state machine (FSM). A FSM is a quintuple $(\Sigma, S, s_0, \delta, F)$, where:

- Σ is the input alphabet (a finite and non-empty).
- S is a finite, non-empty set of states.
- s_0 is an initial state, an element of S.
- δ is the state-transition function that returns a set of transition probabilities: $\delta : S \times \Sigma \rightarrow P(S)$.
- F is the set of final states, a subset of S.

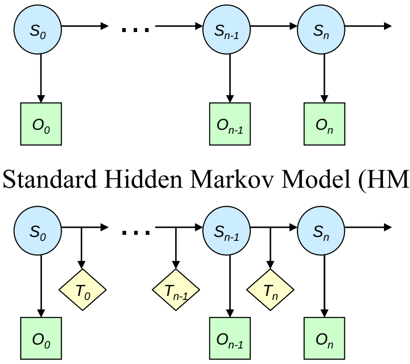
For training,

- Σ is the set of stimuli or trigger events
- S is the set of possible behaviors, i.e. states of the participants.
- s_0 is an initial state.

- δ is the reaction to a stimulus. δ contains both the correct reactions to stimuli defined in a TTP and incorrect reactions that need to be avoided.
- *F* is the end state of a training exercise.

For a training system, states *S* can only be perceived through sensor observations, *O*. Then, behavior analysis is to estimate states $S=\{s_0,s_1,..,s_n\}$ given sensor observation $O=\{o_0,o_1,..,o_n\}$. In our system, the sensor inputs include positions of all participants, their head, body and gun poses and shot/hit data.

From the first look, behavior analysis can be solved using the standard Hidden Markov Model (HMM). However, given the special circumstance of the training application, behavior analysis for training can be modeled using an Augmented HMM (AHMM).



(a) Standard Hidden Markov Model (HMM)

(b) Augmented Hidden Markov Model (AHMM)

Figure 2. HMM and AHMM.

As shown in Figure 2(a), using a standard HMM, a state s_i can be estimated from the observation o_i and the transition probability, $P(S_i|S_{i-1})$, from the previous state s_{i-1}. Therefore, given the estimate of s_{i-1}, $\hat{s}_{i-1}$,

$$\hat{s}_i = \arg\max\big(p(o_i \mid s_i)\, p(s_i \mid \hat{s}_{i-1})\big) \qquad (1)$$

In standard HMM, the transition probability is estimated from training samples or pre-determined. However, in training application, some trigger events are also observable. Therefore, we augment the HMM as shown in Figure 2(b) and $\hat{s}_i$ can be computed as

$$\hat{s}_i = \arg\max\big(p(o_i \mid s_i)\, p(s_i \mid \hat{s}_{i-1}, t_i)\big) \qquad (2)$$

where the data term $p(o_i \mid s_i)$ captures how likely a state is from the observed sensor data, such as position, pose and shots fired; and the generalized prior $p(s_i \mid \hat{s}_i, t_i)$ represent the likelihood of the current state given the previous state and the trigger event. When the trigger event t_i can not be observed, (2) becomes (1), the standard HMM applies. However, when the trigger event t_i is observed, the transition probability will be determined not only by the state estimate at *i-1* but also the trigger event at *i*.

By incorporating trigger events, i.e. stimuli, in the behavior analysis framework, AHMM is particularly suited for training applications and for the need of automated performance evaluation. For example, a platoon is patrolling when they receive a sniper fire. Therefore, the state of the platoon is changed from *patrol* to *reaction_to_sniper*. To estimate the correct state using standard HMM is difficult because we have to design classifiers to correctly recognize the state of *reaction_to_sniper*, which is complex and how a team executes this state can vary significantly. Additionally, what about, in an extreme case, the team does not react to the sniper fire and goes on patrol as usual. Although this is totally wrong, since the HMM recognizes the team's state as *patrol*, the set of performance metrics for *patrol* would be used and this team would not receive a low score for not reacting to the sniper fire.

Different from HMM, AHMM will handle the above scenario correctly. Since the trigger event, sniper fire, is used for computing the transition probability, any state other than *reaction_to_sniper* will receive a zero transition probability, even if data term based on the sensor observation has a high probability that the trainees' state is *patrol*, only *reaction_to_sniper* will be computed as the state of the platoon. Therefore, the performance metrics for *reaction_to_sniper*, the correct performance metrics will be used to evaluate the team's performance. Only when trigger events can not be observed, such as verbal commands, AHMM will rely solely on sensor data to estimate the trainees' states.

Behavior Analysis Algorithms

To estimate the state that trainees are in using AHMM, one needs to define and compute the conditional probability $P(O_i|S_i)$ for all possible states and the prior probability $P(S_i|S_{i-1},T_i)$ for all possible transitions.

$P(S_i|S_{i-1},T_i)$ is generally easy to define. Given a trigger event, all possible transitions and their likelihood will be defined by the TTP and stored in the Training Ontology. Typical trigger events are entering / leaving a zone, being in proximity of a person, a vehicle, a site or an object, receiving fire, explosion, etc.

For simple states, such as walking, running, shooting, the conditional probability $P(O_i|S_i)$ can be manually defined. However, complex activities, such as group formations and group interaction, are difficult to detect using a rule based approach.

We propose a novel feature, Histogram of Oriented Occurrences (HO2) [Cheng, 2008], to model and recognize complex group activities. HO2 captures the interactions of all entities of interests in terms of configurations over space and time and can be used with standard classifiers, such as SVM (Support Vector Machine) for complex activity detection and classification. The output the these classifiers will be normalized as the conditional probability $P(O_i|S_i)$.

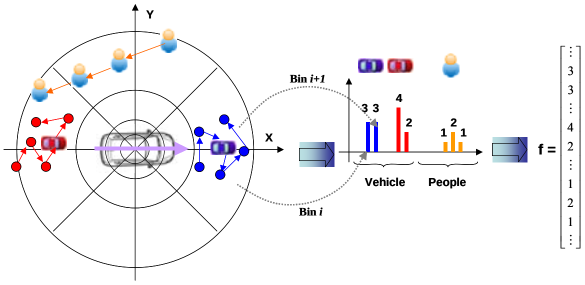


Figure 3. HO2 computation. The reference entity is the middle vehicle. The histograms of vehicle and people occurrences are in the middle.

Figure 3 illustrates how the HO2 feature is computed. Given a reference entity/location, a local coordinate system is centered on and oriented with respect to the reference, and partitions the 2D ground plane into bins. A HO2 feature is a histogram which consists of the number of occurrences for each entity class of interest in each bin. Typical entity classes of interest include Marines, OPFORS, vehicles, buildings, doors and roads. HO2 features can be computed at either a time instance or over a time interval. When computed over a time interval, HO2 is the accumulated occurrences of entities. If the reference entity moves, such as a moving vehicle or person, the partitioning of the ground plane (or the bins) moves with the reference entity. As a result, HO2 captures entity interactions within the frame of reference for the reference entity.

To build one specific event detector, annotated samples of the event are used as the training set. The HO2 features are computed for all entities. Then, an SVM classier is built using this training set.

Training Ontology

The training ontology captures knowledge related to a set of training objectives including TTP, training scenarios and performance metrics. All states and trigger events form the taxonomy in our training ontology. For each state, we also store associated attributes including classifier and the performance metrics for the state. For each state and a given trigger event, the ontology also captures all states that it can transition to.

To develop a general purpose methodology that can be applied to a wide range of training objectives, our system addresses the following two critical needs:

1. Encode TTP and exercises using a formalism for describing concepts and relationships.
2. Create a representation for transitions or trigger events and define their relationships to each state, so that a behavior analysis engine may be easily used.

We address the first need by creating a taxonomy of concepts, including states and trigger events culled from TTP. This taxonomy is used for search and retrieval of relevant concepts from the domain. Our training taxonomy is divided into two sub-hierarchies – a set of concepts representing states (nouns) and a set representing trigger events (verbs). Using Protégé [Noy, 2001], we assign a node to each state, along with the corresponding definition. Similarly, we assign a node to each trigger event and its definition. We show a fragment of this taxonomy below for two states and corresponding taxonomic relationships, "HasType", "HasPart", and "HasSubState".

An ontology fragment for an outdoor MOUT exercise (States/Relationships) contains:

1. Patrolling (State)
HasType (Relationship):
a. Reconnaissance patrolling
b. Raid patrolling
HasPart (Relationship):
Formation
Type (Relationship):
1. Single line column
2. Staggered column
3. Wedge
2. Reacting to Sniper (State)
HasSubState (Relationship):
a. Seeking cover
b. Suppressing sniper
c. Manuevering
d. Blocking escape route
e. Assaulting sniper

For each state in the concept taxonomy, we also associate properties such as "context", "gun pose", etc. Similarly, we link specific performance measures to each concept as associated properties.

To encode the causal relationship among states, trigger events and the logic of an exercise, we adopt the finite state machine (FSM) model. FSM is modular and promotes re-use. We encode a state as a sequence of one or more sub-states, and sub-states may be correspondingly aggregated to form higher-level states. The details of this are presented in the next section on event detection.

We use RDF (Resource Description Framework), a metadata standard that has been recommended by the W3 Consortium to address the need for representing transitions or trigger events. There are a number of advantages to using RDF. First, it provides a way to encode TTP as computer interpretable text. Second, it enables the creation of RDF graphs independently from ontology construction. These graphs encode an independent set of relationships and properties that are associated with the trigger events. These graphs can then be incorporated into the ontology by simply linking them (as references) to the appropriate concept nodes. Third, RDF graphs can be plugged in with the finite state machine formalism in a seamless way [Melnik, 2000], thereby providing input to the behavior analysis computational framework.

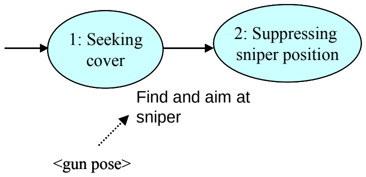


Figure 4. First two sub-states and transition in "Reacting to Sniper" FSM

Figure 4 shows a portion of the finite state machine for the "Reacting to Sniper" scenario. Each circle represents a state and an arrow between states represents a transition as the result of a trigger event. Each trigger event can have one or more properties. Examples of these properties include "gun pose", "context", etc.

As in [Melnik'2000], in Figure 5, we present the RDF encoding of the transition. The transition is centered on the RDF resource, *node*$_i$. We show that this node has four properties:

1. Origin state
2. Destination state
3. Transition (trigger event)
4. Gun pose

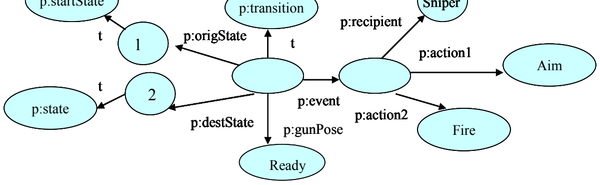


Figure 5. RDF encoding of transition in Figure Y.

Properties associated with the finite-state model are prefixed with "p:", while "t:" represents a RDF type property.

In order to represent trigger events at different levels of granularity in the taxonomy, we decompose events into component actions. For example, a trigger event such as "enter" or "leave" can be decomposed to the primitive action of "walk". Actions that cannot be decomposed further are referred to as primitive actions. Primitive actions are entities that can be detected and classified directly from sensor data. Examples of other primitive actions include 'stop', 'shoot', and 'carry'. Entities at each level of the event hierarchy are mutually disjoint. Entities at each level other than the lowest level (primitive actions) are composed of a time ordered sequence of entities from lower levels. We note that each trigger event and its component actions have associated classifiers that are part of event detection and will be described in the next section.

AUTOMATED PERFORMANCE EVALUATION

For performance evaluation, our system computes performance metrics associated with each state during a training exercise. We also store them in the training information database for AAR. For MOUT training, the following performance metrics are computed:

- **360 degrees Security:** The percentage of a full 360 degrees that is either covered by a Warfighter's weapon or is blocked by a cover.

- **Blocking:** The fraction of the time that all danger spots were blocked by the warfighters, i.e. at least one warfighter points his weapon at each of the danger spots. The danger spot may be a possible sniper position or an approaching vehicle, etc. We use "Aim Margin" to determine the blocking accuracy which needs to be achieved.

- **Cover:** The fraction of time that all warfighters maintain cover. The source of cover can be natural objects such as trees, ravines, hollows, reverse slopes, etc. or man-made such as vehicles, trenches, and craters." [USMC, 2006]. The Warfighters are maintaining cover if the minimum distance of each warfighter from any of the source of cover against the threat direction is below the 'Cover Margin'. The sources of cover are computed from the 3D-model of the training environment. We use Hausdorff distance as the distance measure.

- **Flagging/Muzzling:** A warfighter points his weapon at a friendly. The Flagging score is the total number of detected flaggings.
- **Dispersion Measure:** The average nearest-neighbor distance (NND) per unit time for the warfighter team. Depending on the context, dispersion measure can be useful for signaling “bunching”. For instance, “bunching” may be preferred in an urban context when a unit approaches the corner of a building; it is dangerous in open spaces in a rural context where the entire unit may be exposed.

SENSOR SYSTEM

Figure 6 shows the flow diagram of our data capture system. The central Device Server is responsible for collecting telemetry data from the pan/tilt/zoom (PTZ) cameras. As shown in Figure 7, each warfighter is equipped with an array of sensors including GPS for outdoor and RFID for indoor, Inertial Navigation Sensor (INS), trigger sensors and laser engagement system such as MILES or DITS. Data collected from these sensors are wirelessly transmitted to the central Device Server.

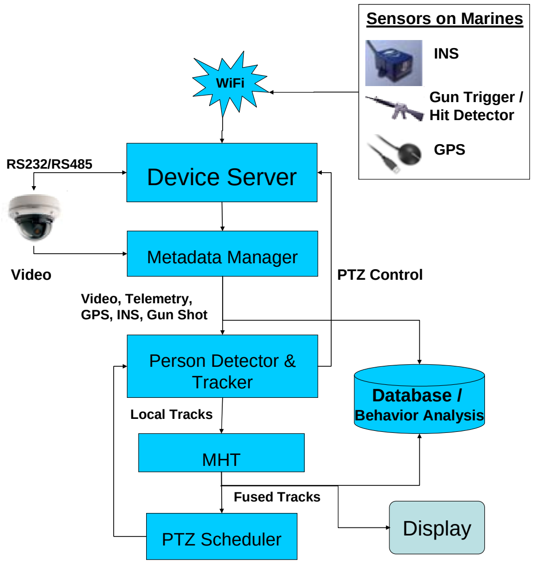


Figure 6. Data Capture and GPS Assisted Active Tracking System Diagram.

From the sensor system, we will compute

- Participants’ position. For indoor training, we use RFID which can achieve about 1 ft accuracy. For outdoor, we use GPS assisted visual track that will be discussed in details in the following section.
- Head and gun orientations are computed from INS data.
- Shot and hit information are captured using a trigger sensor and a laser engagement system.
- Warfighter body poses are estimated from videos.

Active Tracking System

For outdoor training, in order to cover a large area using minimal number of cameras, we use Pan/Tilt/Zoom (PTZ) cameras actively following the trainees during a training exercise.

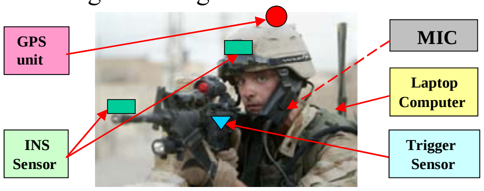


Figure 7. Sensors carried by a war fighter.

As shown in Figure 6, the Device and Video Servers relay data to the Tracker Nodes as well as the database for storage as raw data. Each Tracker Node performs GPS assisted visual tracking and pose estimation of warfighters in a designated camera/view generating local tracks at each node. The local tracks from each Tracker Node are relayed to the Multiple Hypothesis Tracking (MHT) module to be fused into global tracks. The global tracks are stored in the database and sent to the PTZ Scheduler, which together with the PTZ telemetry data computes the appropriate adjustments in the camera pan, tilt, zoom to maintain targets in the field of view.

GPS Assisted Visual Tracking

Although GPS is effective for outdoor tracking, its accuracy is not sufficient for behavior analysis and performance evaluation. Visual tracking using videos on the other hand can provide accurate localization but are not robust against occlusion, view changes and lighting etc. We have developed a GPS assisted visual tracking algorithm combining their strengths.

Our visual tracking algorithm consists of a people detector that uses a set of templates to create people detection likelihoods in a video frame. This process is shown in Figure 8. A set of the person’s silhouette and motion templates are produced using a variety of human walking postures. The mean templates from different walking postures at varying scales (see Figure 8(a)) are saved a-priori. For each video frame, motion blobs are obtained using frame-to-frame registration. These, together with gradient images, are used for correlation-based human template matching to produce the person detection likelihoods as depicted in Figure

8(b). The likelihoods are projected onto a geo-spatial reference plane also used by the GPS data. A local maxima search is used to obtain person localization hypotheses reflecting a high degree of consensus between both the visual detection and GPS localizations for all participants. Tracking over time is then formulated as a correspondence problem between previous track locations and current detection hypotheses, which is solved using bi-partite graph matching. We model GPS error as a Gaussian process and estimate the GPS drift. This improves detection accuracy and provides better track prediction in the case of catastrophic failure of visual tracking.

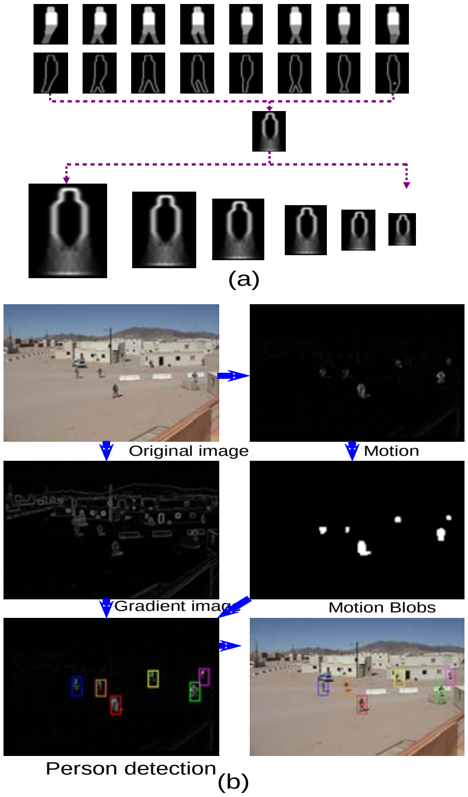


Figure 8: Person Detection and Tracking. (a) Human templates used for correlation based matching. (b) Person detection and tracking steps.

We compute the expected (E) or average localization error, the 50 percentile error and 95 percentile (2dRMS) as shown in Table 1. 2dRMS is a commonly used metric of GPS accuracy.

Table 1: Tracking performance.

	GPS Only	GPS + Video
Median/ 50 Percentile error	4.2 m	0.3 m
2dRMS/ 90 Percentile error	8.0 m	2.4 m
Avg. Localization Error (E)	4.4 m	0.6 m

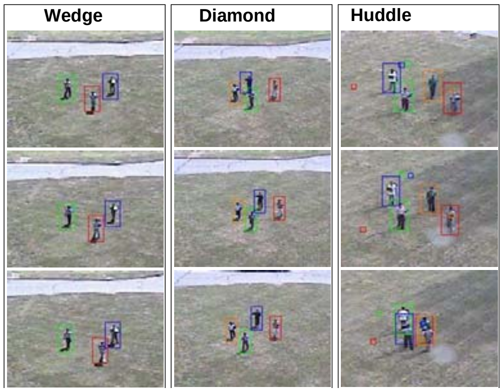


Figure 9: Examples of GPS assisted visual tracking results on various different formations.

Marine Posture Estimation

Using video frames, we also developed Posture Estimation to determine the overall body posture, including standing, kneeling and prone positions, in addition to the direction that the torso, head and weapon are facing (i.e. left, right, towards or away from the camera). A Viola-Jones style multi-class detector [Viola, 2001] for overall appearance is combined with another boosted detector that re-uses object parts for multi-class efficiency [Torralba, 2001] determines the eight upright classes [Wachs, 2009]. The prone postures are distinguished through the aspect ratio of the detected person’s bounding box using a Bayesian classification. The results are smoothed temporally and combined with a non-maxima suppression method using the mean-shift algorithm.

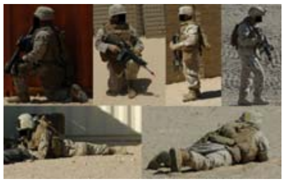


Figure 10: Three stances in two orientations each.

The head position and the head gaze direction are determined with an identical approach after the search region-of-interested is limited by stance-specific priors. The results on preliminary experiments show a detection rate of 80% with a 5% false positive rate. A greater than 50% overlap of the detected with the actual head bounding box is considered a match. The weapon is detected and its orientation is determined using template matching and edge-orientation classification.

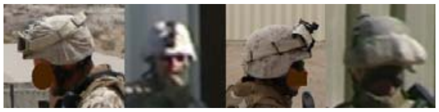


Figure 11: The four detected head orientations.

VISUALIZATION AND AAR

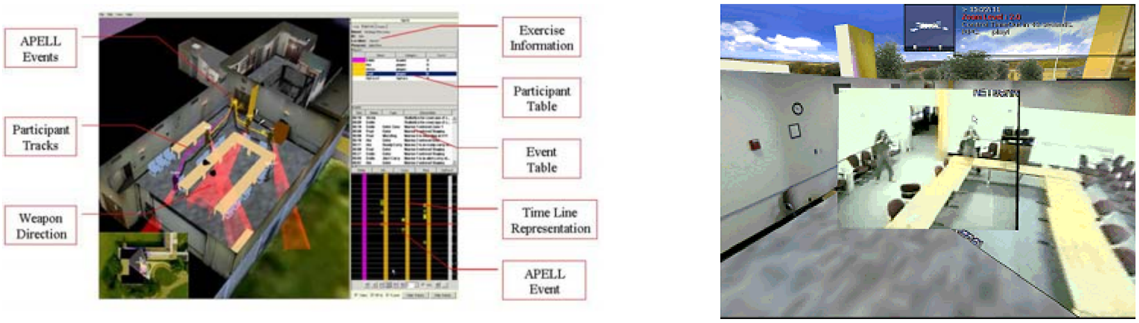
To improve training effectiveness, a training system must allow trainees and instructors to quickly access information, such as lessons learned collected during an exercise for After-Action-Review. We provide a suite of visualization tools that allow a user to view not only videos captured during an exercise, but also tracks poses, and events in an interactive and easy-to-use manner. Two displays are provided by our system and they are used simultaneously in a synchronized fashion. They are (1) Symbolic Map Display and (2) Video Flashlight Display.

Symbolic Map Display

In Figure 12(a), we show a snapshot of a typical exercise as it is displayed by the Symbolic Map Display. The left side of the Symbolic Map Display shows a 3D-model of the MOUT environment. Tracks, gun poses, and head poses (not shown) of participants, including warfighters (with ID) and OPFORS are overlaid on the 3D-model. Events detected by the system are marked by symbols on the track. A user can view the tracks from different view angles and can click on any event symbol to view the event description. A user can also turn on/off the gun pose or head view of any participants.

The right side of the display consists of four different parts: (1) General exercise information including exercise ID, starting, ending time and exercise duration are shown at the top. (2) The participant information is shown below the exercise information. (3) The event table consisting of time, participant, the type of the event and description is shown in the middle. (4) A time line representation of the exercise overlaid with events is shown at the lower-right corner.

The Symbolic Map Display allows users to view an exercise at any instant and track movements forward and backward in time. A user can also drag the red time line to any location and play back from there. Additionally, the user can synchronize the Symbolic Map Display with the Video Flashlight Display to view the symbolic representation and video at the same time.



(a) Symbolic Map Display for AAR (b) Video Flashlight Display for AAR

Figure 12. Visualization and AAR

Video Flashlight Display

Our system uses multiple video cameras to cover a MOUT facility and captures the entire exercise. To better view the videos captured by different cameras, Sarnoff has developed a Video Flashlight Display system [Kumar, 2003, Hsu, 2000] that can seamlessly integrate multiple video streams into a unified live display. Each of the videos is projected on the 3D-model of the MOUT environment. The dynamically textured model is rendered and displayed by the Video Flashlight system. By controlling the viewpoint and the view angle, a user can get a birds-eye view of all the activity covering the area of regard or the user can zoom in and focus on an object or person of interest. An example of the Flashlight Video display is show in Figure 12(b), where two video streams are projected onto the 3D-model of the room being cleared by a four-man Marine team.

EXPERIMENTS

The prototype training system has been implemented and used to test mock warfighter Marine exercises including both indoor room clearing and outdoor patrol scenarios. Two of the outdoor training exercises that we have focused on are "React to Sniper" and "Patrolling". Figure 13 shows an illustration of one of the simulated "React to Fire" exercises performed by four Marines divided into teams of two. To perform the exercise correctly first the Marine unit needs to move towards the building to obtain cover. Once in cover one team blocks the sniper with their weapons while the other team moves across to engage and neutralize the sniper threat. We achieve an average detection rate of 85% on our mock exercises.

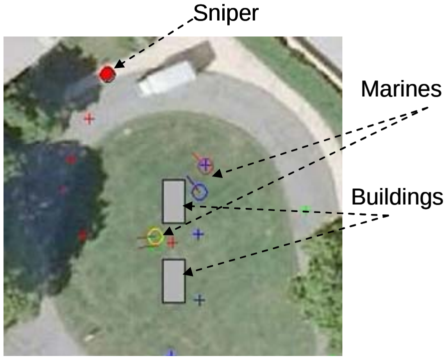


Figure 13: Bird's eye view illustration of a "React to Sniper" exercise.

In Figure 14 we show the metrics computed by our system for the "React to Fire" exercise illustrated in Figure 13. For "Flagging" and "Blocking" we used an "Aim Margin" parameter of 10 degrees. The "Cover Margin" parameter for "Cover Score" was set at 2 meters for this exercise. These values were selected based on empirical testing. From the plots in Figure 14 it can be seen that the Marines maintain a good blocking score through out the exercise. The cover score is high in the time window of 30-55 seconds. This is the time period the warfighters regroup behind the buildings and begin the assault on the sniper. The flagging score is low except for the time period around 70s mark where multiple muzzling acts were detected. An interesting metric to notice is the NND score which hovers between 1.5 and 3 meters. This metric shows that the warfighters did not disperse and maintained cohesiveness within the teams.

In Figure 15, we show a muzzling event during an indoor room clearing exercise. The muzzling occurred across the extent of the room and not observable in any one camera. In Figure 15(a), the muzzling is clearly visible in the Symbolic Map Display and is detected by the automated event detection algorithm. In Figure 15(b) and 15(c), we show the video frames containing the two participants involved in the muzzling. The participant marked by the red circle in Figure 15(b) accidentally pointed his weapon at the participant marked by the red circle in Figure 15(c). Since the two participants are far from each other, without our system, it would be very difficult to spot these kinds of muzzling events.

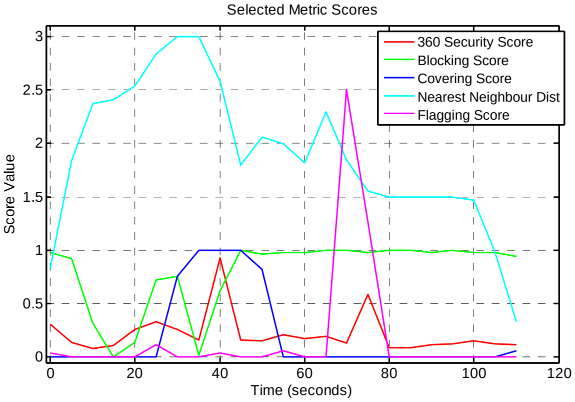


Figure 14: Performance Metric for a "React to Sniper" Exercise.

CONLUSION

We have developed a computational framework for automated behavior analysis and performance evaluation that effectively incorporates TTP and designed training scenarios. We propose a novel Augmented Hidden Markov Model (AHMM) for behavior analysis that takes advantage of the controlled training environment and planned stimuli for a training exercise. A novel HO2 feature is developed to capture complex interactions among multiple players in a group activity. To capture trainee behavior, the prototype training system captures and computes tracks, poses and actions of the participants and automatically assesses the performance of warfighters using a training ontology. This prototype system was used for simulated Marine Corps exercises. The results show that the prototype system can accurately detect participants' states, mistakes, such as muzzling, automatically. The detected events and computed performance metrics provide power tools for advanced AAR capabilities.

The proposed training system can significantly reduce instructor's workload of setting-up, controlling and monitoring an exercise and evaluating trainees' performance. Learning points including mistakes made and performance trends are automated computed and indexed in a database that can be quickly retrieved for an AAR. The multi-media AAR combining both video and iconic map display improves the effectiveness and efficiency. With real-time behavior analysis and performance evaluation, our system enables closed-loop training where the training scenarios and difficulty levels are automatically adjusted in real-time based on the performance and needs of the war fighters. In summary, such a training system will allow trainees to learn more and will allow instructors to teach more in a shorter period of time.

In future work, we plan to expand our MOUT training ontology to capture not only the expert knowledge of MOUT operations, including procedure and strategies of warfighters, but also the behavior and strategies of OPFORs and civilians. The ontology may be used both for performance evaluation and for control of the behaviors of OPFOR and civilians in the exercise. Finally, the entire training system could easily be modified and extended by changing the ontology to cover a larger range of MOUT training exercises.

REFERENCES

Cheng, H., Yang, C., Han, F., Sawhney, H. (2008). HO2: A new feature for multi-agent event detection and recognition. *Computer Vision Pattern Recognition Workshop*, pp.1-8.

S. Hsu, S. Samarasekera, R. Kumar, H. S. Sawhney (2000), Pose Estimation, Model Refinement, and Enhanced Visualization Using Video, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Hilton Head Is., SC, vol. I, 488-495.

S. Jung, Y. Guo, H. Sawhney, R. Kumar. (2008). “Action Video Retrieval Based on Atomic Action Vocabulary,” Proc. ACM Int’l Conf. .on Multimedia Information Retrieval, Vancouver, British Columbia.

R. Kumar, S. Samarasekera, A. Arpa, M. Aggarwal, V. Paragano, K. Hanna, H. Sawhney, M. Sartor (2003), Monitoring Urban Sites using Video Flashlight and Analysis System, *GOMAC Proceedings*, Tampa Florida.

Robert J. Fontana (2004). Recent System Applications of Short-Pulse Ultra-Wideband (UWB) Technology. *IEEE Transaction on Microwave Theory and Techniques, vol. 52 no. 9, September 2004, pp. 2087-2104.*

N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. Fergersen, M. A. Musen (2001). Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, 16, 2, 60-71, 2001.

S. Melnik, H. Garcia-Molina, A. Papepcke (2000). A Mediation Infrastructure, for Digital Library Services. *ACM Digital Libraries*, 123-132.

Viola, P. & Jones, M. (2001). Robust Real-time Object Detection, *2nd Intl Workshop on Statistical and Comp. theories of Vision*, Vancouver.

Wachs, J.P., Goshorn, D., & Kölsch, M. (2009). Recognizing Human Postures and Poses in Monocular Still Images. *Intl. Conf. on Image Processing, Computer Vision, and Pattern Recognition (IPCV).*

Torralba, S.A., Murphy K.P., and Freeman W.T. (2007). Sharing visual features for multiclass and multiview object detection, *IEEE PAMI*, 29:5, pp. 854-869.

Camouflage, Cover and Concealment, Lesson Plan. USMC, Weapons and Field Training Battalion. January 26, 2006.

T. Zhao, M. Aggarwal, R. Kumar and H.S. Sawhney (2005), Real-time Wide Area Multi-camera Stereo Tracking, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA.*

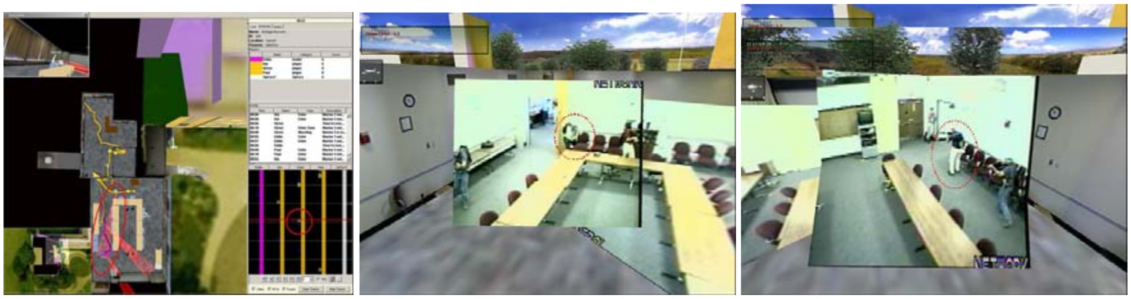


Figure 15. Muzzling / Flagging detected, though the participants involved across the room from each other. (a) Muzzling is marked in red. (b) Video frame shown the participant marked in the red circle pointing his gun at the participant in Figure 15(c) marked in the red circle. Since the two participants are across the room, no a single video frame captures both of them during the muzzling.