Juan P. Wachs

# A Training and Assessment Tool for Warfighter Tasks

## 1. Statement of the problem

In the Navy, there is a constant need to improve the skills of the warfighter while lowering training times [1]. Well-trained, highly skilled warfighters will experience fewer mishaps, which is a top priority as stated by former Secretary Rumsfeld's memorandum of June 22, 2006: [2] "Fund as a first priority those technologies and devices that will save lives and equipment". Our approach to this problem is the development of innovative intelligent tools, which can monitor the warfighter's progress and guide him to the desired state of preparedness, taking into consideration different learning styles and physical-physiological factors.

**Goal 1:** We propose a system based on computer vision to capture expert knowledge and skill in the performance of tasks, which can be used to evaluate a trainee's skills. This system will lead to cost savings in personnel training programs as well as providing a safe and effective testing environment for the warfighter's skills.

**Goal 2:** We will work to develop robust computer vision algorithms integrating multi-views and multi-modal vision techniques for 3D tracking.

## 2. Background and relevance to previous work:

Computer assisted training tools have been shown to accelerate the learning process [3]. The main requirements of such tools are that they: a) identify optimal strategies for task completion; b) assess and diagnose task performance in specific domains; c) increase the efficiency and completion time of the training process; and d) reduce training costs. Task fulfillment can be assessed in multiple ways. For example, some tools assess the performance of the task according to the accuracy of the result (i.e., tabulation device at a shooting range). Here we focus on the process itself; this means the correctness of the sequence of actions.

Human skill can be defined as the learned power of doing something competently, or the ability of a person to use acquired knowledge effectively, so that her performance of a task is maximized [3]. A task is composed of human actions, including hand and body movements. Automated performance measurement technology should have the capability to “infer” tasks by capturing human movements. Computer vision is one approach towards human movement segmentation and analysis. For example, in [4] the problem of 3D human motion modeling using monocular vision and articulated models is addressed. Human posture tracking and classification using stereo vision is studied in [5]. In [6], a multi-perspective and multimodal video based system for 3D body tracking is researched. Each of these techniques has relevance toward the training system envisioned here.

The utilization of computer vision human movement segmentation and analysis is the first step towards a computer-based skill training, which has several advantages over traditional skill training methods [3]: a) traditional training takes a long time, since each instructor is responsible for several trainees; b) the models used in traditional training lack realism, while the use of real devices may be costly and risky (i.e., dismantling a bomb); c) in traditional training, the evaluation of the trainee is based on the subjective evaluation of the expert, rather than on quantitative performance measures; and d) computer-based training captures warfighter expertise, which can reduce the manpower load of expert trainers. The two first obstacles have been successfully addressed through the introduction of Augmented Reality (AR) systems [7]. AR systems can be used without supervision and provide realistic case scenarios.

Still, the warfighter performance evaluation issue has not been fully addressed by intelligent systems and the following challenges remain: a) to improve sensing techniques to capture the relevant features of human-body movement sequences, and b) to devise a computational model for representing and assessing the warfighter's skills. The introduction of hand and body gesture recognition systems [8] can provide robust techniques to cope with these challenges. The warfighter's gestures involved in the task (i.e., dismantling a bomb, assembling a weapon) can be segmented. Therefore, a learning system can be trained. Similarities (likelihood) of the beginner's gestures to the modeled correct gestures can serve as a measure of the task performance.  Some examples of hand gesture recognition systems are [9]-[14].

In our system we plan to address the problem of tracking by using a combination of cutting edge approaches (such as articulated body tracking, gesture recognition, stereo vision, multi-spectral imaging, monocular vision-based SLAM, and MoCap sensors) to obtain 3D human motion data. In addition, we propose to use hand and body movements in a combined approach to extend the scope to analysis of human behaviors, which has applications in surveillance and "asset protection".  A training system based on these tools will enhance the combat capability of US warfighters.

## 3. General Methodology

The main goal of this research is to obtain observations of the experienced warfighter's continuous hand and body movements while performing an assembly task and then modeling the skills demonstrated in the observations as a reference for the evaluation of trainees. More specifically, let us assume an assembly task which we want to use as a case study for skill training. For this task, it is possible to define one or more correct sequences of actions (strategies) involved in the task. A sequence of actions will be obtained from the trainee, and his performance will be reflected by the similarity of his sequence to one of the correct sequences existing in the model's memory for that task. To aid the trainee in improving her skills, the system will provide her with the correct sequence closest to the sequence of movements that she chose. The problem is to find the prototype sequence from all the sequences stored in the model, such that the performance skill of the trainee is maximized.

Our methodology differs from previous works [3] in that: (a) we consider that there may be more than one correct body movement sequence to perform a task; (b) our system will provide the trainee with an evaluation and will show the trainee the closest strategy (in previous works only an evaluation grade was supplied); and (c) our system will consider multiple views and multi-modal vision techniques for 3D tracking in real-time of the human body.

These goals can be achieved by breaking the problem into functional modules: (a) human movement observation and feature acquisition, (b) model development and training, (c) skill grading and suggestion for improvement, and (d) validation of the model.

a. Movement observation and feature acquisition: video observations of experienced tutors will be obtained, each performing an assembly task several times. As an example, a task may include the following actions {"insert", "pull", "hold", "release", "rotate", "move to the side"}. To track in real-time the tutor's arm and body movements during the task, encumbered and unencumbered approaches will be considered.

For the encumbered approach, MoCap sensors that will supply 3D information about each part of the body may be attached to the warfighter's body. Since there are many points involved in human motion, "interesting" points [15] will be selected as representatives of skills demonstrated in assembly tasks. For the unencumbered approach, vision-based SLAM, articulated body tracking, and mono and stereo vision may be considered initially to obtain position information. Further on, we are interested on combining multi-perspective (i.e., network camera views) and multi-modal (i.e., thermal-infrared and color) video based systems for robust and real-time 3D shape tracking. These approaches will yield spatial and temporal information of the body part tracked.

As an example, if the region of interest is the hand, the position of each hand can be described as the x, y and z components of the segment connecting a reference point and the hands. The motion sense can be captured by the subtraction of consecutive frames, and the relative displacement of the hands. Then the vertical and horizontal components of the velocity and acceleration of each hand can be determined. If a higher level of description is required, a posture of the hand can be represented by attaching fiducials to the fingertips, and the center of the palm can be used as a reference point.

This motion information is incorporated in the feature vector $X(k_t)$ at the time t, and normalized such that all the parameters of the feature vector are equally scaled. Using this method, local temporal features are created. As a side note, the successful selection of the features will highly determine the accuracy of the posture classification [16].

b. Model development and training: Different approaches based on pattern recognition algorithms [17] will be considered to model the skill behind temporal observations. To include the stochastic nature of the learning process, Hidden Markov Models (HMM) [18] constitute a promising approach. The temporal model will be trained with the feature vectors obtained from the sequence of observations (i.e., the feature vectors) from one or more subjects, performing the same task, using the same sequence of actions for several trials. Once the model is trained (for example a HMM), the score of a trainee can be measured by some metric function (such as the log probability) applied to the new observation sequence. Since we allow a task to be performed in different sequences, we create a family of these models, each trained with a given sequence of observations. We call every sequence used to train each model a prototype sequence. Each model will be trained with the same number of "tutors," or experts, in an assembly task.

c. Skill grading and suggestion for improvement: After the family of models is trained, the trainee will feed the system with an observation sequence. This sequence will be compared to the prototype sequences, and a score vector representing the similarity of the trainee's sequence to each prototype sequence will be determined. The highest value of the score vector represents the best matched prototype sequence to the trainee's sequence. The highest score will be the trainee's task performance grade, and the best matched prototype will be suggested to the user for further practice. This score is expected to increase, reflecting the trainee's performance improvement, see Fig. 1.

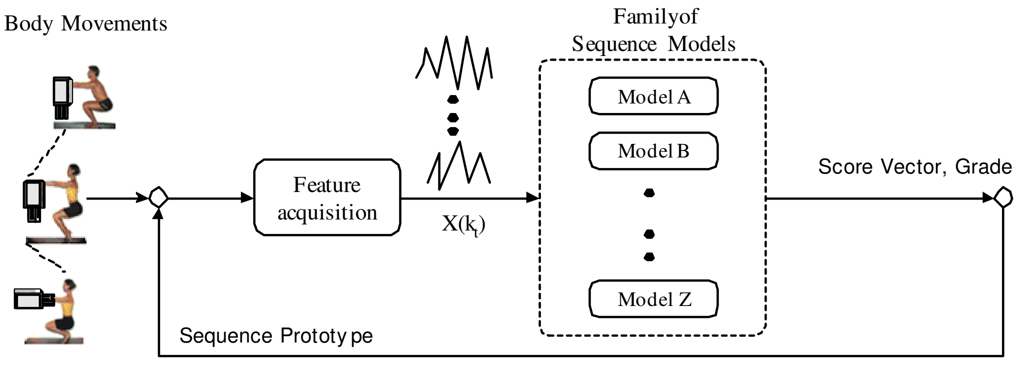

**Fig. 1. Flow chart diagram of the modeling and assessment of human body movement**

d. Validation of the model: To validate the approach, two experiments will be performed: a) comparison of different sequences for the same skill; and b) skill evaluation as a function of time. For the first experiment, two sequences for the same task will be considered (i.e. model A and B). Each sequence represents 2 different valid strategies to perform the task. The samples for the two models will be collected from several experienced trainers. The number of samples for the training and testing set can be obtained using k-fold cross validation [19], to improve the statistical significance of the results. For example, given 15 samples for training the system, each time 14 samples can be used for training and 1 for testing. Hence a total of 15 training "sessions" can be conducted. The average evaluation grade for all the sessions will be used to test the reliability of the model to the actual skills of the trainees.

In the second experiment, a different set of samples collected from tutors will be used, where 2/3 of them will be used for training and the rest for testing. Additionally, samples in temporal order will be collected as a test set from non-experienced subjects. A statistical t-test using the average across the sessions will be used to compare the scores obtained when testing the system with tutors and with beginners.

## 4. New or unusual technique

A real-time efficient method will be developed based on the optimal integration/selection of multi-perspective (i.e., several camera views) multi-modal (i.e., thermal infrared and color) video based system for robust tracking of 3D body regions. Most approaches only use a single view from either color or thermal infrared cameras. This research will use an integrative approach that will give more robust results.

## 5. Expected results, significance and application

From the initial experiment suggested, using the data collected from only the tutors, we expect to find that the grade for an observation sequence similar to A to fit the model of the prototype sequence A, will be higher than the grade for this observation to fit the model B, and that the results will be statistically significant.

Regarding the second experiment, we hope to see that a higher evaluation score will be obtained when testing the model with the tutor than when tested with the beginner. We also hope to find that the score will rise according to the temporal order of the sample observations of the beginner, which will demonstrate that the trainee learns the skill tested.

Possible causes of low correlations and/or the rejection of our statistical hypothesis may be caused by the reduced number of tasks available in a laboratory environment. This can be alleviated by additional virtual "realistic" tasks obtained when incorporating haptic techniques, immersive displays and augmented reality. The success of the experiments will show that the motions of an expert soldier used in assembly tasks are extremely efficient and then can be adapted to an automatic skill training system.

We expect to implement the experimental platform suggested in the proposal to a real-time diagnostic assessment prototype for Knowledge, Skills and Abilities (KSAs) area at the Naval Postgraduate School in Monterey. Such a platform can carry multiple implications: a) the enabling of accelerated training progression by identifying warfighters' individual potentials; b) efficient real-time methods for modeling and task assessment in specific domains (KSA); c) techniques for comparison and monitoring of the initial stage and the desired end stage of the trainee warfighter (with corrections suggested as necessary).

The suggested approach for body and hand movements modeling and evaluation which is strongly related to the field of “Modeling Activity and Understanding Behavior” has broader applications such as people tracking from surveillance cameras for asset protection and security.

In summary, the main contribution of the methodology suggested is a reconfigurable platform based on integration of machine vision modalities with broad applicability to security/defense and skill training applications. A computer assisted warfighter’s training and skill assessment was selected for proof of concept. This can be applied to all stages in warfighter development. Initially Marine recruits could be selected based partially on skills assessment. For advanced Marines the diagnostic assessment tool will be used for continuous training. The deployment of this tool will drastically change the way the Navy does training. The accuracy and precision that this method will lend to motor skills will increase the effectiveness of warfighters performing assembly tasks (and later other types of tasks) and therefore the constant preparedness of the Navy. In addition, each warfighter will have a personalized training program, increasing the efficiency of the training budget and its cost effectiveness. The eventual decrease in necessity of manpower for training will free the trainers to do what they do best. This goes in hand with the mission at the MOVES institute and the Naval Postgraduate School of enhancing the operational effectiveness of the US forces by providing superior training in the field of modeling and simulation.

## 5. References

[1] SECDEF Report to Congress. 1998. Actions to Accelerate the Movement of the New Workforce Vision. *Section 912(d) Questions 10, 13, and 14.*

[2] SECDEF Memo June 22, 2006. Reducing Preventable Accidents.

[3] Chen J, Yeasin M, Sharma R. 2003. Visual modeling and evaluation of surgical skill. *Pattern analysis and applications*. 6(1): 1-11.

[4] A. D. Sappa, N. Aifanti,S. Malassiotis and M. G. Strintzis. 2004. 3D gait estimation from monoscopic video. *International Conference on Image Processing ICIP '04*. 3:1963-1966.

[5] S. Pellegrini, L. Iocchi. 2006. Human Posture Tracking and Classification through Stereo Vision. In *Proc. of Intern. Conf. on Computer Vision Theory and Applicartions.*

[6] S. Y. Cheng, S. Park, M. M. Trivedi. 2005. Multi-Perspective Thermal IR and Video Arrays for 3D Body Tracking and Driver Activity Analysis. *IEEE International Conference on Computer Vision and Pattern Recognition.*

[7] M. A. Livingston, J. E. Swan II, S. J. Julier, Y. Baillot, D. Brown, L. J. Rosenblum, J. L. Gabbard, T. H. Höllerer, and D. Hix. Evaluating system capabilities and user performance in the battlefield augmented reality system. In *Proc. NIST/DARPA Workshop on Performance Metrics for Intelligent Systems*, Gaithersburg, MD, Aug. 24-26 2004.

[8] T. E. Murphy, C. M. Vignes, D. D. Yuh. and A. M. Okamura. 2003. Automatic Motion Recognition and Skill Evaluation for Dynamic Tasks. *Eurohaptics*.

[9] J. Triesch and C.V.D. Malsburg. 1998. A Gesture Interface for Human-Robot Interaction. *Proc. of 3th IEEE Intl. Conf. on Automatic Face and Gesture Recognition*. 546-551.

[10] T. S. Huang, and V. I. Pavlovic. 1995. Hand Gesture Modeling, Analysis, and Synthesis. *Proc. of Intl. Conf. on Automatic Face and Gesture Recognition*.

[11] K. Abe, H. Saito, and S. Ozaw. 2002. Virtual 3-D interface system via hand motion recognition from two cameras. *IEEE Trans. Systems, Man and Cybernetics, Part A*, 32(4):536-540.

[12] X. Yin and M. Xie. 2003. Estimation of the fundamental matrix from uncalibrated stereo hand images for 3D hand gesture recognition. *Pattern Recognition*. 36(3): 567-584.

[13] W. L. Seong. 2006. Automatic Gesture Recognition for Intelligent Human-Robot Interaction in *Proc. of the 7th Intl Conf on Automatic Face and Gesture Recognition* (FGR06). 645 – 650.

[14] M. Kölsch and M. Turk. 2005 .Hand Tracking with Flocks of Features. *In Video Proc. CVPR IEEE Conference on Computer Vision and Pattern Recognition*.

[15] D. O. Olguín, and A. Pentland. 2006. Human Activity Recognition: Accuracy across Common Locations for Wearable Sensors. *IEEE 10th International Symposium on Wearable Computing*.

[16] J. Wachs, H. Stern, Y. Edan. 2005. Cluster Labeling and Parameter Estimation for the Automated Setup of a Hand-Gesture Recognition System. *IEEE Transactions on Systems, Man and Cybernetics. Part A*. 35(6): 932-944.

[17] M. Nechyba and Y. Xu. 1995. Human Skill Transfer: Neural Networks as Learners and Teachers. *International Conference on Intelligent Robots and Systems, Human Robot Interaction and Cooperative Robots*. 3. 314-319.

[18] L. R. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*. 77 (2). 257–286.

[19] E. M. Tzanakou, Supervised and Unsupervised Pattern Recognition: Feature Extraction and Computational Intelligence, Rutgers University, Piscataway, New Jersey, USA, pp. 50-52,1999.