The Limitations of Static Performance Metrics for Dynamic Tasks Learned Through Observation

Amy E. Henninger
Soar Technology, Inc.
317 N. First St.
Ann Arbor, MI 43031
amy@soartech.com

Avelino J. Gonzalez
Michael Georgipoulos
Ronald F. DeMara
School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816-2450
ajg,mge,rfd@isl.engr.ucf.edu

Keywords:
Performance Metrics,
Learning By Observation,
Human Skill Representation

ABSTRACT: *A recent report developed by the National Research Council (NRC) for the Defense Modeling and Simulation Office (DMSO) encourages the use of real world, war-gaming, and laboratory data in support of the development and validation of human behavioral models for military simulations. This paper reviews existing validation metrics used in human behavioral modeling exercises and discusses the limitations of these metrics. Common to the metrics examined is the fact that they have been applied to a specific type of human behavioral model, a low-level, reactive skill model. These models, in turn, are related by the fact that they have been developed through some form of learning by observation. Thus, the scope of this paper is constrained to reviewing fidelity metrics of low-level, reactive skill models that have been created from observational data. To this end, it is assumed that model fidelity is correlated with similarity to true human performance.*

This paper will be designed around two metrics found in the literature on skill models developed through learning by observation and it will give detailed illustrations showing where these metrics are deficient. It is anticipated that through the explanation of where these metrics fall short, improved metrics can be developed. Concepts for improving these metrics and candidate metrics are presented, but a completely functional alternative has not yet been investigated.

1. Introduction

A recent report developed by the National Research Council (NRC) for the Defense Modeling and Simulation Office (DMSO) encourages the use of real world, war-gaming, and laboratory data in support of the development and validation of human behavioral models for military simulations [1]. This paper reviews validation metrics that a have been used for a specific type of human behavioral model, a low-level, reactive skill model. To better understand the metrics reviewed, the models to which they correspond are also reviewed. Common to these models is the use of learning by observation as the underlying framework for model development. Thus, the scope of the metrics reviewed in this paper is limited specifically to metrics of low-level, reactive skill models that have been created from observational data. To this end, it is assumed that model fidelity is correlated with similarity to true human performance. In other words, a model that is more similar to the data source from which it was generated would be of higher fidelity than a model that is less similar.

"Learning by Observation" is a term typically associated with the Artificial Intelligence/Machine Learning (AI/ML) communities. AI/ML are relatively young, multi-disciplinary research fields concerned with developing computational theories of learning and constructing machines with "learning" capabilities. They embrace principles of disciplines such as bayesian methods, computational complexity theory, control theory, information theory, philosophy, psychology and neurobiology, and statistics. Their immaturity coupled with their diversity gives rise to a vernacular that is not without variation. For purposes of clarity, this section presents the interpretations used in this paper for the terms "reactive skill" and "Learning by Observation".

1.1 Reactive Skill

The term "behavior" can be defined as any observable action or reaction of a living organism [2]. Some psychologists would also include conscious phenomena such as cognitions, perceptions, and judgments. This extension complements the way the term "behavior" is used within the military simulation community as well as in this paper. For instance, a CGF might represent the selection of a reactive behavior by means of a decision table. This construct maps situational awareness information (perceptions) to reaction (judgment). Of course, cognitive phenomena like these are not directly observable, but they can be inferred from low-level behaviors (e.g., firing weapons, driving, etc) that are observable. These types of "low level", observable behaviors, on the other hand, are often referred to in the literature as "skills" [3] or "human control strategies" [4]. In [4] human skills are grouped into two categories: 1) action skills and 2) reaction skills, where the former are defined as being open-loop (e.g., drop-kicking a ball) and latter are characterized as requiring feedback to a human in a control loop (e.g., driving). Low-level or skill-level behaviors used in the CGF or military simulation community (e.g., driving, scanning, etc.) would correspond to the "reaction skills category" as a series of low-level decisions blended with some display of motor skills. Of course, "action skills", as defined by [4] would also be evidenced in the military simulation domain. For example, this term could be applied to a behavior like a tank's main gun firing or a dismounted infantry (DI) unit falling prone. The effort described in this paper focuses on the former type of skill-level behaviors, "reaction skills" and refers to them simply as skill-level behaviors that may be represented by human skill models or human control models.

1.2 Learning by Observation

While the term "behavior" can be defined in somewhat vague terms, the AI/ML community has tolerated small variations in the definitions because the underlying idea is familiar and hence, these definitions can generally be understood. But, another term that is central to this paper, "Learning by Observation", has a less familiar definition. For example, some authors [5][6] use this phrase to suggest that the data being used to develop the learning model are simply acquired through means of observation as opposed to introspective methods. Other authors, however, associate the phrase "Learning by Observation" specifically with unsupervised learning [7]. This interpretation suggests that Learning by Observation can actually occur through data/knowledge acquisition techniques other than "observation". Because this latter camp of authors defines the phrase "Learning by Observation" as being synonymous with unsupervised learning, it is restricted to those forms of learning where there is no a priori classification of observations into sets exemplifying desired concepts. That is, this definition doesn't account for forms of supervised learning that use observational data. This document uses the phrase "Learning by Observation" to indicate that the data must be acquired by observational means and that the learning model may be formed through either supervised or unsupervised learning techniques. Observational data are non-experimental data, based simply on observing behavior without trying to manipulate it experimentally.

2. Metrics for Human Control Models

A number of researchers have attempted to model human driving or flying skills (e.g., acceleration, steering, vehicle following, etc) in an effort to develop robotic and or simulated drivers [8] [9] [10] [11] [12] [13]. Some researchers in the robotics domain above have been successful in testing their systems in real-world applications. For example, in the development of Autonomous Land Vehicle In a Neural Network (ALVINN), [8] was able to test model performance by letting the robotic-controlled vehicle drive itself at speeds up to 55 mph for distances of over 90 miles on public roads. These vehicles have proven themselves capable of driving both during the day and night, driving on a variety of roads under adverse weather conditions, avoiding obstacles, and even performing parallel parking.

Researchers developing these models for a simulated domain typically rely on a subjective evaluation of the model's performance in the simulated environment. A lack of metrics for models like these in the CGF community has been recognized by a variety of authors [14] [15] [16]. [14] recognizes the multi-dimensionality of the decision space and the non-linearity of the response surface as being a major part of the challenge. In [15] the importance of incorporating a temporal measure into a metric is acknowledged, and [16] stresses the point that the choices and decisions people think they make is not always consistent with the choices and decisions they actually make. The authors of this paper maintain that one part of what makes this task so monumental is the fact that there exist a myriad of different types of models and representations for different types and levels of decisions, skills, and tasks. In light of this, it is reasonable to anticipate the need for a combination of different types of metrics. After all, prevalent views suggest that no single modeling technique can adequately represent all types of human behavior. It would follow then, that more than one type of metric would be needed to measure all types of human behavior models. As such, this paper is scoped to specifically consider one type of metric, a metric for low-level reactive skill models that have been created from observational data.

The following subsections: 2.1, 2.2, and 2.3 discuss examples of these types of low-level, reactive skill models, present the metrics used to measure the models' fidelity, and explain the limitations of these metrics. Then, in section 3, concepts for improving the metrics are proposed and a framework for applying the metrics to CGFs is discussed.

2.1 Lane Switching Model and Metric

[11] and [12] developed a neural network architecture using recurrent back-propagation to mimic human performance in simple lane switching and acceleration tasks. The basic configuration of this system can be seen in Figure 1.

The subject's task in this system was to drive his car around the track, switching lanes and adjusting speed as necessary to avoid a collision. The perspective of the presentation is such that the controlled car is always at 3 o'clock and the remaining traffic moves relative to its position. Data were collected from the subject driving the track with mouse controls, and the final network consisted of 42 inputs, two hidden layers, and 2 output representing the speed and the lane. The goal of the network was to emulate the subject's performance and style.

Researchers concluded that experiment was successful, but acknowledged that this conclusion was subjectively based. However, they also offered some quantitative measures based on static averages of their model inputs compared with averages based on source data. While the attempt to provide an objective performance metric was commendable, the metric on which they based their conclusions was limited. For example, in an attempt to objectively validate a simulated model of human driving, they compare averages of important values (e.g., "Distance to Traffic in Front of Control Vehicle at the Time the Control Vehicle Switches Lanes") of the true data and the model's execution data. However, the power of this metric is limited because 1) it is a static approximation to a dynamic model and 2) it ignores the interactive effects between this and other variables in the system. Figures 2 and 3 serve to illustrate these points with hypothetical data. For purposes of illustration, assume the blue plot (human driver) in Figure 2 is the true distribution of "Distance to Traffic in Front of Control Vehicle at the Time the Control Vehicle Switches Lanes" over lane switching events. Similarly, the red plot also represents the "Distance to Traffic in Front of Control Vehicle at the Time the Control Vehicle Switches Lanes", but it represents the model as the controller not the actual human driver.

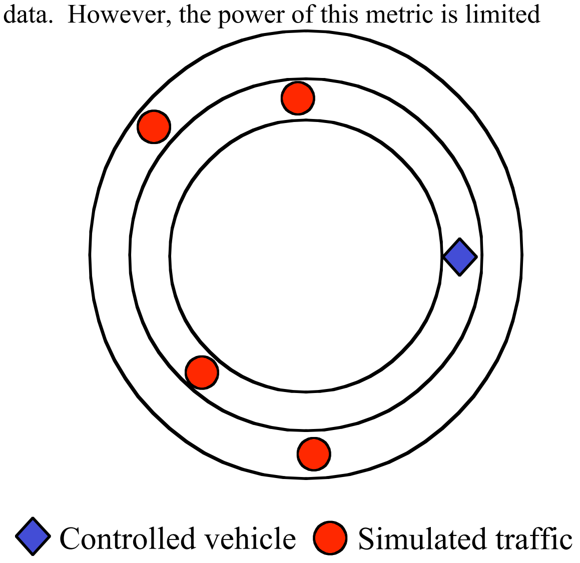

Figure 1. Fix and Armstrong (1990) System Presentation

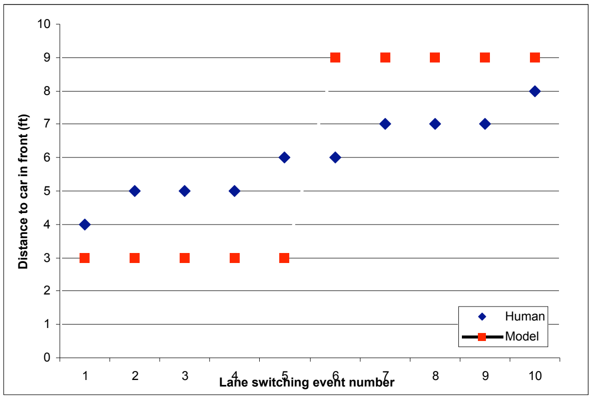

Figure 2. Distribution of "Distance to Traffic in Front of Control Vehicle at the time the Control Vehicle Switches Lanes" Ordered Over Lane Switching Events

The mean of the human lane switching data is 6 feet, the same as the mean for the model lane switching data. Thus, using this metric, a very good or almost perfect approximation is suggested. However, Figure 2 clearly demonstrates that this is not a valid interpretation. That is, even though both distributions have a mean of 6, they're not identical. This measure is misleading because it fails to take into account the distribution of those individual data with respect to other predictive dimensions. This concept is further illustrated in Figure 3, which shows one possible relationship between the "Distance to Traffic in Front of Control Vehicle at the Time the Control Vehicle Switches Lanes" and the controlled vehicle's speed.

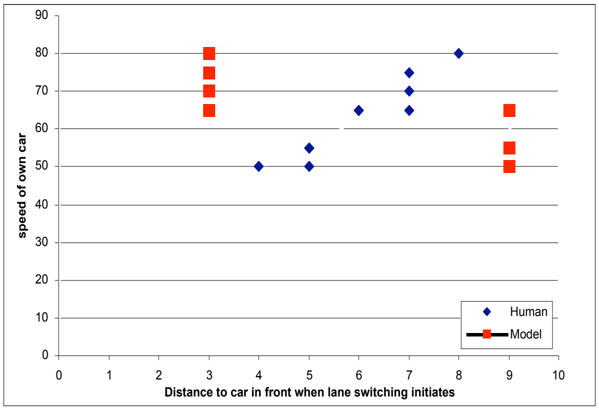

Figure 3. Distribution of "Distance to Traffic in Front of Control Vehicle at the Time the Control Vehicle Switches Lanes" vs Controlled Vehicle's Speed

Figure 3 shows an inverse relationship between the "Distance to Traffic in Front of Control Vehicle at the Time the Control Vehicle Switches Lanes" and the controlled car's speed for the human performance data and the model performance data. That is, while the human generally switched lanes at closer distances when it was moving slower and at farther distances when it was moving faster, the model's performance did the complete opposite. Thus, even if the model's data distribution in Figure 2 was deemed acceptable for application, Figure 3 reveals yet another potential limitation of this metric, its inability to measure fidelity in light of other influential variables with which it may co-vary.

2.2 F-16 Flying Model and Metric

[10] used an artificial neural network, the Artificial Neural System for the Representation and Collection of ACM Decision-Making Expertise (ARCADE), to model air combat tactics for adversary aircraft (F-16). ARCADE creates a mapping from the tactical environment (inputs) to the appropriate maneuver (output), where a maneuver is defined as a change in the current flight path of the aircraft. For example, a change in any one or any combination of factors in the input vector (e.g., blue aircraft angle of attack, blue altitude, blue and red airspeed, blue and red pitch orientation, etc.) may produce a change in the red F-16's maneuver (e.g., pitch, thrust, etc).

The input vector for the ARCADE network repeated the same set of parameters at 5 different points in time, T_{now}, T_{now-1}, T_{now-2}, T_{now-5}, and T_{now-10}. This enabled ARCADE to assess temporal information in determining its mapping. The output vector provided a maneuver recommendation in terms of desired pitch, desired angle, and commanded velocity; in a manner which would enable its recommendation to be translated into a set of parameters that could drive the operation of the aerodynamics model. The size of the output time window was also provided to the dynamics model to define how quickly the flight condition variables should be changed.

The ARCADE system used the Back-propagation training algorithm with (at least) a 3 layered, feed-forward architecture. An initial scaled-down ARCADE prototype system was developed to model a simple Lead Pursuit/Intercept algorithm. In other words, instead of modeling human behavior, the initial prototype attempted to reproduce algorithmic behavior. This prototype consisted of approximately 1500 input/output vector pairs and required approximately 250,000 iterations through the training set to converge.

Validation of the model was performed by integrating the model into software that regulated the path updates of a simulated entity, and then visually observing its behavior. According to the author, the network's pursuit behavior mapped very closely to the algorithm's profile in most cases. In addition to this subjective evaluation of the model's performance, Crowe offered the mean square error (MSE) derived from training as second form of measurement used to assess model fidelity.

Mean Squared Error is a value used to express the difference between the true data and the model's approximation to that data over the range of the data. Specifically, the MSE is given by equation 1, where d represents the desired output, y represents the actual output, and p represents the pattern number.

$$E^{p} = \frac{1}{N}\sum_{p=1}^{N}\left[d(p) - y(p)\right]^{2} \qquad (1)$$

Since the gradient descent procedure employed by the Back-propagation training algorithm attempts to minimize the MSE, a common method for measuring the performance of a back-propagation network is to calculate the MSE over the entire set of training data and/or validation data. Thus, a lower MSE would indicate a better performing model. While this could be an adequate measure in a less-dynamic, deterministic domain, it is severely limited in a dynamic, stochastic domain such as modeling a human control model. First, it shares the same problem as that explained in section 2.1. For a stochastic process, a static error criterion (such as MSE) does not consider the distribution of the error over the source data. Moreover, for a dynamic process, model errors can feed back on themselves to produce trajectories that are not characteristic of the source data or are even potentially unstable. Again, the problem lies within the distribution of the error over the source data. Thus, a static error measure, such as MSE, does not provide sufficiently satisfactory model validation for a dynamic process; and for stochastic systems, one cannot expect equivalent trajectories for the source data and the learned model, given the same initial conditions.

2.3 M1A2 Driving Model and Metric

In [17], the authors model the near-term driving behaviors from data collected from a subject matter expert interacting with a simulated scenario in a table-top M1A2 driver's station (see Figure 4). In this system, they had two, independent networks, one to predict the change in the driver's heading and one to predict the change in the driver's speed. Each of these networks was trained according to the back-propagation algorithm and each used a feed-forward network configured with five inputs, five hidden nodes, and one output.

Figure 4. Driver at M1A2 Simulator

In this study, the driving model was being used as a synchronization model for a dead-reckoning application. In other words, instead of being part of a controller or generator of behavior for a simulated entity, these models were being used to predict human behavior as part of a DIS dead-reckoning model (i.e., the neural networks replaced the Newtonian models). Because of the uniqueness of the application, an application specific metric (i.e., reduction in ESPDUs) was adopted to measure the models' performance. However, this metric was application specific. If, for example, those models were being evaluated as controllers or generators of a simulated entity's behavior, then [17] would have also lacked a meaningful metric. For example, Figure 5 shows the trajectory and speed data over the three runs made by the SME. The trajectory plot is presented relative to the route's center-line, and both the trajectory and speed plots denote the point at which the way point change occurs. Alternatively, Figure 7 shows the trajectory and speed data over the runs made with the neural networks developed with the SME's source data. This representation was compared to the currently used nearterm movement model in ModSAF [18] (see Figure 6), a prevalent SAF system used in the military training community. While it's clear from a visual comparison that the neural network was far more similar to true human performance than the ModSAF nearterm movement model was, no quantitative performance metric exists in the CGF community to measure the strength of this similarity.

Because there is no meaningful metric with which to measure this effort, there is no methodical means of evaluating and improving current modeling strategies. Moreover, there is no methodical means of comparing the performance of alternative modeling strategies. Clearly, if the CGF community is to advance the state-of-the-art in human behavior representation, it needs a metric that would facilitate the process of improving models and allow for the comparison of alternative modeling techniques.

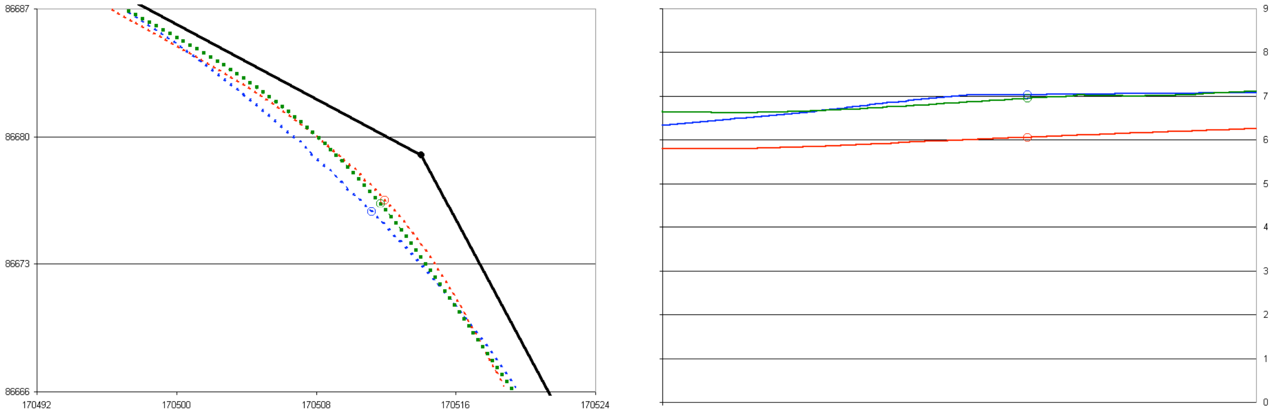

Figure 5. Bunker6's Trajectory (left) and Speed (right, m/s) Data for Runs 1, 2, and 3

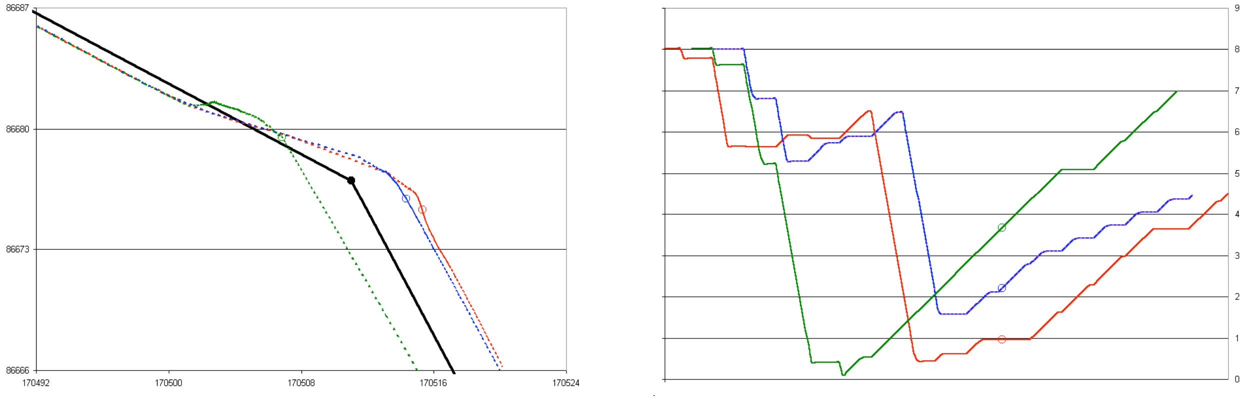

Figure 6. ModSAF's Trajectory (left) and Speed (right, m/s) Data for Runs 1, 2, and 3

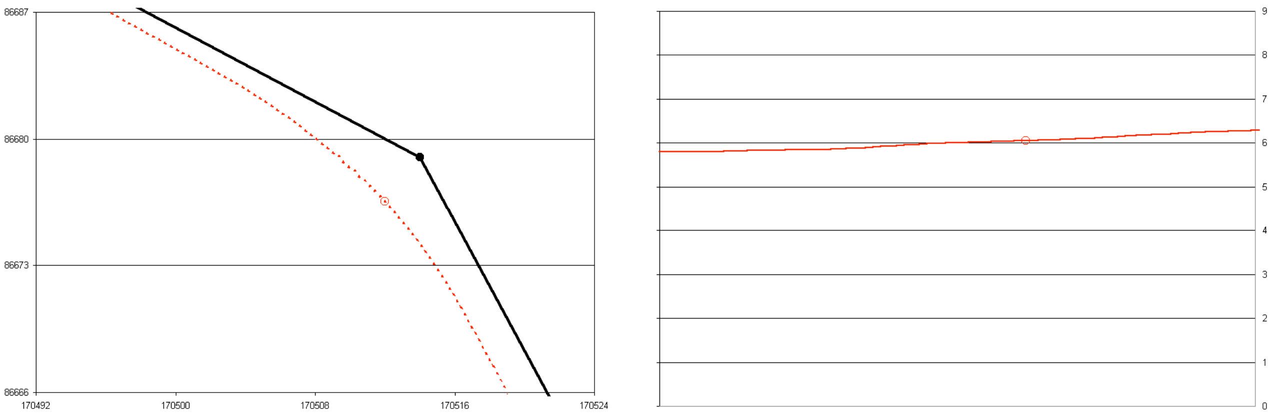

Figure 7. Neural Network's Trajectory (left) and Speed (right, m/s) Data for Runs 1, 2, and 3

Route centerline ▬▬▬ **Vehicle path** ········ **Waypoint** ○

3. Candidate Metric

Recently, researchers at the Robotics Institute at Carnegie Mellon University [4] have developed a metric that accounts for the problems discussed in the previous section. That is, this metric is able to analyze the multi-dimensional control trajectories generated by the human as well as the corresponding learned model. [4] accomplishes this by correlating observations of the model data to a Hidden Markov Model developed from the source data. A HMM is a stochastic state transition construct that can be completely defined by the following triplet: $\lambda = \{A, B, \pi\}$, where A represents the probabilistic $n \times n$ state transition matrix, B represents the $l \times n$ output probability matrix with l discrete output symbols, and π represents the *n*-length initial state probability distribution vector. For some observation sequence, O , the probability that a given observation sequence is generated from the model λ (i.e., $P(O|\lambda)$) can be evaluated. Since two HMMs are defined to be equivalent iff:

$$P(O|\lambda_1) = P(O|\lambda_2) \tag{2}$$

a similarity measure can be derived by 1) characterizing the source trajectories by training a "validating HMM", and 2) cross evaluating that HMM with the model's trajectory. Thus, the final measure becomes a ratio of the two:

$$\sigma = \frac{P(O^V, T^V, \lambda^*)}{P(O^*, T^*, \lambda^*)} = \frac{P_2}{P_1}, \sigma > 0 \tag{3}$$

where λ^* = HMM trained on observation sequence O^* of length T^*
O^V = observation sequence of length T^V, which we wish to compare to O^* through λ^*

Prior to calculating these values, each dimension of the data must be normalized and converted to a sequence of vectors, and then these vectors must be discretized through some through some vector quantization process. These techniques have been applied by Nechyba and well documented in the speech recognition and signal processing communities.

The use of a metric like this would meet the challenge issued by National Research Council by using human performance data to validate models. Further, though the applications considered in this paper were low-level reactive skill models developed through learning by observation, it is expected that this type of metric could generalize to other situations. For example, instead of human performance data, any set of "truth" or "perfect" data could be used as the verifying set. Thus, similarity metrics for models could be derived against any data set deemed as the "gold-standard". Also, this metric should be able to support any type of dynamic process. Thus, one would expect it to generalize to "higher" levels of behavior other than skill-level tasks.

The lack of an objective, meaningful fidelity metric for human behavioral models in military simulations has impeded progress in evaluating the performance of current modeling strategies as well as in comparing the performance of alternative modeling strategies. This paper proposes a candidate fidelity metric that could fill such a gap.

4. References

[1] Pew, R.W., and Mavor, A.S., eds. (1998). Modeling Human and Organizational Behavior: Application to Military Simulations. Washington, DC: National Academy Press.

[2] Hilgard, E.R., Atkinson, R.L., and Atkinson, R.C. (1979). Introduction to Psychology, Seventh Edition. Harcourt Brace Jovanovich, Inc. New York, NY.

[3] Lee, S., and Chen, H. (1996). Robot Skill Discovery Based on Observed Data. Proceedings of the 1996 IEEE International Conference on Robotics and Automation, Minneapolis, MN. pp. 2694-2699.

[4] Nechyba, M.C. and Xu, Y. (1997). Learning and Transfer of Human Real-Time Control Strategies. Journal of Advanced Computational Intelligence, vol.1, pp. 137-154.

[5] Gonzalez, A., Georgiopoulos, M., DeMara, R., Henninger, A., and Gerber, W., (1998). Automating the CGF Model Development and Refinement Process by Observing Expert Behavior in a Simulation, The Eighth Conference on Computer Generated Forces and Behavioral Representation Proceedings. Orlando, FL., May, 1998.

[6] Van Lent, M. and Laird, J. (1998). Learning by Observation in a Tactical Air Combat Domain, In Proceedings of the Eighth Conference on Computer Generated Forces and Behavior Representation. Orlando, FL., May, 1998.

[7] Michalski, R.S., Carbonell, J.G., and Mitchell, T.M. (1986). Machine Learning: An Artificial Intelligence Approach Volume II, Los Altos, CA.: M. Kaufmann Publishers.

[8] Pomerlau, D., Thorpe, C., Longer, D., Rosenblatt, J.K., and Sukthankar, R., (1994). AVCS Research at Carnegie Mellon University. Proceedings Of Intelligent Vehicle Highway Systems America 1994 Annual Meeting, p. 257-262.

[9] Pentland, A. and Liu, A., (1995). Toward Augmented Control Systems. Proceedings of Intelligent Vehicles, vol. 1, page 350-55.

[10] Crowe, M. (1990). The Application of Artificial Neural Systems to the Training of Air Combat Decision-Making Skills. Proceedings of the 12th Interservice/Industry Training Systems Conference, Orlando, FL. Nov., 1990.

[11] Fix, E., and Armstrong, H.G., (1990). Modeling Human Performance with Neural Networks. Proceedings of the International Joint Conference on Neural Networks, vol. 1, pages 247-252.

[12] Fix, E., and Armstrong, H.G., (1990b). Neural Network Based Human Performance Modeling. In Proceedings of IEEE National Aerospace and Electronics Conference, vol. 3, pp. 1162-5.

[13] Henninger, A., Gonzalez, A., and Georgiopoulos, M., (2000). Modeling Semi-Automated Forces with Neural Networks: Performance Improvement through a Modular Approach, The Ninth Conference on Computer Generated Forces and Behavioral Representation Proceedings. Orlando, FL., May, 2000.

[14] Harmon, S.Y., and Youngblood, S.M., (1999). Validation of Human Behavior Representations. In Proceedings of the Spring Simulation Interoperability Workshop, Orlando, FL., March, 1999.

[15] Gonzalez, A.J., and Murillo, M,. (1999). Validation of Human Behavioral Models. In Proceedings of the Spring Simulation Interoperability Workshop, Orlando, FL., March, 1999.

[16] Velt, C.T. (1993). Developing Validated Behavioral Representations for Computer Simulations. In Proceedings of the 3rd Conference on Computer Generated Forces and Behavior Representation. Orlando, FL., March, 1993, pp. 417-428.

[17] Henninger, A., Gonzalez, A., Gerber, W., , Georgiopoulos, M., and DeMara, R., (2000). On the Fidelity of SAFs: Can Performance Data Help, Proceedings of the '00 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC). Orlando, FL., November, 2000.

[18] Smith, J. (1994). Near-term Movement Control in ModSAF. In Proceedings of the Fourth Conference

in Computer Generated Forces and Behavior Representation. Orlando, FL, May 1994.

Author Biographies

AMY HENNINGER is a Senior Scientist at Soar Technology, Inc., an Ann Arbor based company specializing in the representation of intelligent, automated computer generated forces. Recently, Dr. Henninger completed the requirements for a doctoral degree in computer engineering at the University of Central Florida (UCF). She's a former Research Fellow for the Army Research Institute at U.S. Army STRICOM. She has earned B.S. degrees in Psychology, Industrial Engineering, and Mathematics from Southern Illinois University, an M.S. in Engineering Management from Florida Institute of Technology, and an M.S. in Computer Engineering from UCF.

AVELINO GONZALEZ received his bachelor's and master's degrees in Electrical Engineering from the University of Miami, in 1973 and 1974, respectively. He obtained his Ph.D. degree from the University of Pittsburgh in 1979, also in Electrical Engineering. He is currently a professor in the School of Electrical and Computer Science at UCF, specializing in human behavior representation.

MICHAEL GEORGIOPOULOS is an Associate Professor in the School of Electrical and Computer Science at UCF. His research interests lie in the areas of neural networks, fuzzy logic and genetic algorithms and the applications of these technologies in cognitive modeling, signal processing and electromagnetics. He has published over a hundred papers in scientific journals and conferences.

RONALD DEMARA is a full-time faculty member in the School of Electrical Engineering and Computer Science at UCF. Dr. DeMara received the B.S.E.E.. degree from Lehigh University in 1987, the M.S.E.E. degree from the University of Maryland, College Park in 1989, and the Ph.D. degree in Computer Engineering from the University of Southern California, Los Angeles in 1992. His research interests are in the areas of Parallel and Distributed Computing, Networking, and Simulation.