Dorneich, Mathan, S., Ververs, P.M, and M.C., Whitlow, S.D. (2007), "An Evaluation of Real-time Cognitive State Classification in a Harsh Operational Environment", *Proceedings of the Human Factors and Ergonomics Society*, Baltimore, MD, Oct 1-5.

AN EVALUATION OF REAL-TIME COGNITIVE STATE CLASSIFICATION IN A HARSH OPERATIONAL ENVIRONMENT

Michael C. Dorneich, Santosh Mathan, Patricia May Ververs, Stephen D. Whitlow
Honeywell Laboratories
3660 Technology Drive, Minneapolis, MN 55418

This paper describes an evaluation conducted with a full platoon of 32 Soldiers at Aberdeen Proving Grounds' MOUT site in Aberdeen, MD. The objective was to assess the cognitive workload classification techniques driven by neuro-physiological (EEG) and physiological (ECG) sensors. In a first ever evaluation of real-time cognitive monitoring in the harsh operational environment, the assessment culminated in a three phase, 24-hour mission consisting of a coordinated Route Reconnaissance, a Cordon and Search of a village, and a Hasty Defense operation. Task load levels were manipulated by introducing unexpected and unplanned events requiring re-planning and extensive coordination by the leadership (high task load) as well as lulls in the activity in which part missions were executed flawlessly with little variations on the preplanned, well versed drill (low task load). Four leaders (Platoon Leader, Platoon Sergeant, Squad Leader 1, and Squad Leader 2) were equipped with sensors to measure and output cognitive state in real-time. The fused EEG and ECG workload classification approach reached 95% accuracy depending on the individual and the amount of data used to train the classifier. This level of success implies that Augmented Cognition workload assessment tools enable the ability to move beyond subjective workload rating scales, such as NASA TLX and Cooper Harper ratings, to more objective measurements of real-time cognitive state metrics in almost any conceivable operational environment.

INTRODUCTION

Work in the field of Augmented Cognition began by establishing the ability to classify cognitive processing (attention, working memory, executive function, and sensory memory) with laboratory tasks known as Psych 101 tasks. Gradually over the past three years, researchers have moved from the laboratory environment to the field environment, introducing the artifacts (motion, electrical, networking traffic and disconnects) and stressors (information overload, physical load, competition, and threat of pain) inherent in the operational environment to which the technology would be transitioned. This paper describes an assessment of the ability to classify cognitive workload level in an unconstrained, free-play operation with Soldiers executing missions in an urban terrain environment. An evaluation was conducted with a platoon of Soldiers at Aberdeen Proving Grounds' MOUT (Military Operations in Urban Terrain) site in Aberdeen, MD. The objective was to assess the cognitive workload classification techniques driven by neuro-physiological (EEG) and physiological (ECG) sensors. In a first ever evaluation of real-time cognitive monitoring in the harsh operational environment, the assessment culminated in a three phase, 24-hour mission consisting of a coordinated Route Reconnaissance, a Cordon and Search of a village, and a Hasty Defense operation. Task load levels were manipulated. Unexpected and unplanned events required re-planning and extensive coordination by the leadership, and resulted in a high task load period. Low task load periods consisted of lulls in the activity, or missions that were executed flawlessly with little variations on the preplanned, well versed drill. Four leaders (Platoon Leader, Platoon Sergeant, Squad Leader 1, and Squad Leader 2) were equipped with sensors to measure and output cognitive state in real-time.

Realizing the vision of an augmented cognition system in the context of an ambulatory Soldier has been constrained by several challenges. First, as Schmorrow and Kruse (2002) noted, processing and analysis of neurophysiological data have been largely conducted off-line by researchers and practitioners. However, in order for Augmented Cognition technologies to work in practical settings, effective and computationally efficient artifact reduction and signal processing solutions are necessary. Second, inferring the cognitive state of users demands pattern recognition solutions that are robust to noise and the inherent nonstationarity in neurophysiological signals (Popivanov & Mineva, 1999). Third, understanding the fluctuations of cognitive state in applied environments requires the development of means to collect reliable neurophysiological data outside the laboratory. Fourth, experiments must be designed, often under conflicting constraints (e.g. operational realistic tasks vs. well-understood, controlled lab tasks), to effectively evaluate classification accuracy. Finally, compact and robust form factors (e.g., size, weight, ruggedness) associated with neurophysiological sensors and processors are a matter of critical concern.

The use of EEG as the basis for cognitive state assessment was motivated by characteristics such as good temporal resolution, low invasiveness, low cost, and portability. While EEG offers several benefits, there are shortcomings related to the noise artifacts described above and the nonstationarity of the neural signal pattern over time. Despite these challenges, research has shown that EEG activity can be used to assess a variety of cognitive states that affect complex task performance. These include working memory (Gevins & Smith, 2000), alertness (Makeig & Jung, 1995), executive control (Garavan, Ross, Li, & Stein, 2000), and visual information processing (Thorpe, Fize, & Marlot, 1996). These findings point to the potential for using EEG measurements as the basis for driving adaptive systems that demonstrate a high degree of sensitivity and adaptability to human operators in complex task environments.

SYSTEM DESCRIPTION

The system constructed to assess cognitive state classification algorithms consists of 1) sensors to collect raw physiological (ECG) and neuro-physiological (EEG) data, 2) mobile semi-rugged computer platforms to process the raw sensor data into cognitive state classification assessments, 3) a wireless data infrastructure to send the classification assessment of subordinates to leaders, 4) signal processing to process the raw sensor data and remove/flag any compromised data, and 5) real-time cognitive state classification.

Sensors

EEG data were collected from the Advanced Brain Monitoring (ABM) EEG sensor headset (Figure 1 left). Differential EEG were sampled from six bipolar channels CzPOz, FzPOz, F3Cz, F3F4, FzC3, C3C4 at 256 samples per second with a bandpass from 0.5 Hz and 65 Hz (at 3 dB attenuation) obtained digitally with Sigma-Delta A/D converters. Quantification of EEG in real-time was achieved using signal analysis techniques to identify and decontaminate eye blinks, and identify and reject data points contaminated with electromyography (EMG), amplifier saturation, and/or excursions due to movement artifacts (see Berka, 2004).

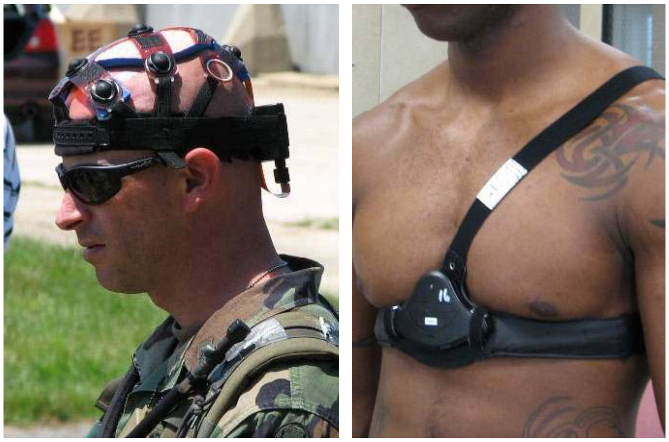

Figure 1. EEG (left) and ECG (right) sensors.

The Hidalgo Vital Signs Detection System (VSDS, see Figure 1 right) measures heart rate, respiration rate, and body motion and position. The evaluation utilized the ECG waveform (2 Views, sampled at 256 Hz) and the three-axis accelerometry waveforms (sampled at 25.6 Hz) signals.

Mobile Processing and Wireless Data Network

Each of the four primary Soldier participants (PL, PSG, SL1, and SL2) was followed by a member of the experimental personnel in the role of "shadower." Each shadower carried a specially designed backpack (based on the MOLLE system) that contained a Panasonic Toughbook CF-51 equipped to receive Bluetooth communication from the subject's EEG, ECG, wireless mic, and head-tracking systems. Each shadower remained within the 30 meter range of their participant to ensure Bluetooth connectivity. Additionally, the shadower wore a Web-cam and logged video to the Toughbook. The participant wore a wireless mic, and the resultant audio stream was multiplexed into the Web-cam video In addition to logging the data, the raw sensor data were processed on the Toughbook using Cognitive State Classification algorithms to produce a real-time assessment of the subject's cognitive state. That cognitive state assessment was then transmitted to the base station via the wireless data network that employed a 900 MHz radio modem system (Figure 2).

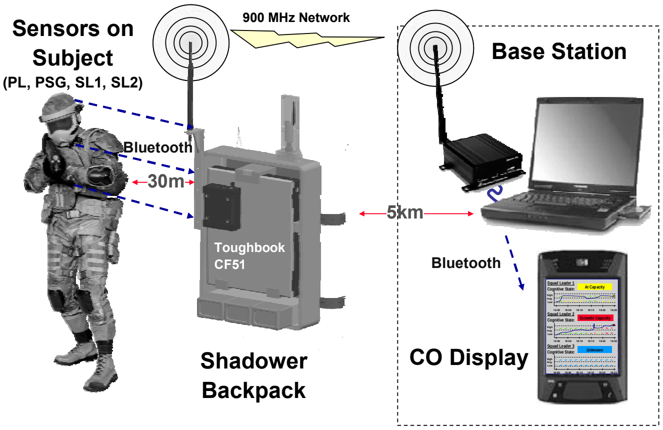

Figure 2. Data is transmitted from sensors (Soldier) to processor (shadower) to the base station.

Signal Processing

Decontaminated EEG was segmented into overlapping 256 data-point windows called overlays. An epoch (the temporal window of analysis) consisted of three consecutive overlays. Fast-Fourier Transform (FFT) was applied to each overlay of the decontaminated EEG signal multiplied by the Kaiser window ($\alpha = 6.0$) to compute the power spectral densities (PSD). The PSD values were adjusted to take into account zero values inserted for artifact contaminated data points. The PSD between 70 and 128 Hz was used to detect EMG artifact. Overlays with excessive EMG artifacts or with fewer than 128 data points were rejected. The remaining overlays were then averaged to derive PSD for each epoch with a 50% overlapping window. Epochs with two or more overlays with EMG or missing data were classified as invalid. For each channel, PSD values were derived for each one-Hz bin from 3 Hz to 40 Hz and the total PSD from 3 to 40 Hz. Relative power variables were also computed for each channel and bin using the formula (total band power/total bin power).

Real-Time Cognitive State Classification

Estimates of spectral power formed the input features to a pattern classification system. The classification system used parametric and nonparametric techniques to assess the likely cognitive state on the basis of spectral features; i.e. estimate *p(cognitive state | spectral features)*. The classification process relied on probability density estimates derived from a set of spectral samples. These spectral samples were gathered in conjunction with tasks that were as close as possible to the eventual task environment.

The classification system utilized a support vector machine (SVM) to discriminate between low and high task load. Support vector machines are linear classifiers that use a quadratic optimization procedure to find an optimal orientation and location for a discriminating hyperplane between two classes. The optimization procedure finds a location and orientation for the hyperplane that lies as far away as possible from examples in each class that are likely to be confused with each other. Separating hyperplanes that are identified using this procedure has been shown to maximize generalization performance (Vapnick, 1999). Although they are linear classifiers, SVMs can be used to solve non-linear problems by means of the so-called kernel trick. Data that may not be linearly separable in the original feature space can be projected into a high dimensional space where the data may be linearly separable. The SVM used in this effort employed a radial basis function kernel with a kernel parameter of 1 and a slack parameter of 0.05.

METHOD

Objective

The principal hypothesis tested was as follows: the Cognitive State Classification algorithms would be able to differentiate periods of high and low cognitive workload using a combination of physiological (ECG) and neuro-physiological (EEG) sensors. Classification analysis focused on how well can the classifier discriminate between workload classes in an inherently noisy and dynamic environment?

Participants

The evaluation utilized a full Platoon if 32 Soldiers from the North Carolina National Guard (NCNG) Combined Arms Battalion, as shown in Figure 3.

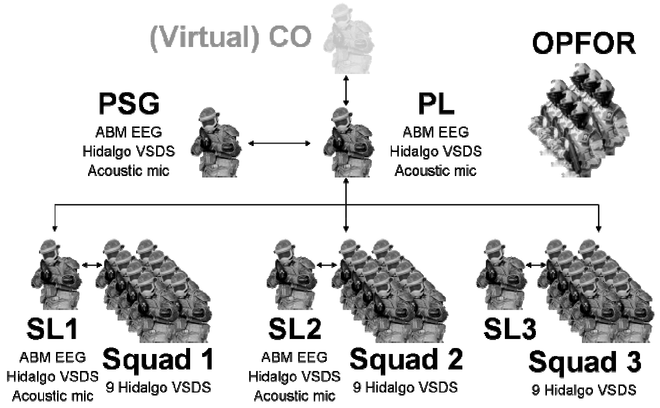

Figure 3. Platoon participants and their equipment.

Data were collected from four participants: the Platoon Leader (PL), the Platoon Sergeant (PSG), the Squad One Leader (SL1), and the Squad Two Leader (SL2). Opposition Forces (OPFOR) were staffed by remaining members of the NCNG. The NCNG Company Commander acted as the CO.

Tasks

There were two principal phases of the 12-day training session where experimental data was. During the period between days 3-10, the platoon conducted part-mission training where they repeated a set of tasks for a 3-4 hour period. The tasks changed each day. The final day of the experiment was a 24-hour full mission training session, divided into three 8-hour phases: 1) conduct dismounted movement along the lines of communication to objective to ensure routes are free of mines and obstacles (during a 5km night march, an IED detonated along their path), 2) conduct a Cordon and Search of the Objective to kill, capture, or expel opposition forces operating in this urban area, as well as to capture and destroy any explosives uncovered, and 3) defend Objective for an extended period and reported any enemy activity in and around this key terrain.

While this evaluation focused primarily on the PL, PSG, SL1, and SL2, the activities of subordinates and responses from senior leaders had a direct impact on their stress levels. There were a host of stressors that the platoon-level training exercise used in the MOUT facility, summarized in Table 1.

Table 1. Stressors in a MOUT environment.

Category	Example Stressors
Loss of sight	Distributed squads
Confusion	Changes in the plans, conditions, and mission; Loss of communications
Realism	Extended operational period (e.g. 24 hours of operation) in the Urban Facility
Fatigue	Extended movement to the facility followed by an assault and then occupation of the site for long periods in a defensive posture
Uncertainty, Threat	Use of OPFOR to prevent friendly forces from gaining control of the Urban Facility to "hit" the friendly forces at different times
Evaluation Stress	Use of simunitions (soap bullets that sting)
Surprise	Impose unexpected elements that affect plan
Severe Weather	Periods of high heat, humidity, intense rainfall

Procedure

The independent variable in the evaluation was workload (all phases). The experimental scenarios were manipulated to ensure definable periods of high and low cognitive workload. Periods of low workload could include completing initial paperwork, reporting activities, pre-planning, long hasty defense position, consolidation/transition, after action, and periods of low activity during missions. High workload periods were characterized by multiple task performance under time pressure and fatigue. Examples of high workload were re-planning due to change in circumstance (e.g. enemy location, available squads, loss of communication, etc.), directing squad movements during pre-assault, squads in assault, managing multiple communications (i.e. responding to commanders, squad leaders, or other platoon leaders), or call for fires/backup. Stressors that contributed to high workload included a degree of frustration or stress, loss of communication, lack of asset availability, and loss of situation awareness of squad locations and activities.

Data Analysis

The metric used to evaluate classification performance is the Area under the Receiver Operating Characteristic (ROC) curve (see Duda, Stork, & Hart, 2001). ROC curves plot true positives (on the y-axis) against false positives (on the x-axis) as a threshold for discriminating between targets and distracters is varied. It is widely used to evaluate human and machine signal detection capabilities. The ROC curve provides a way to assess the degree of overlap between the output of a classifier for two classes of data. Perfect classification produces an area under the curve value (Az) of 1.0, while chance performance produces an Az value of 0.5.

Ground Truth

In order to calculate the accuracy of the classification approach, classifier results are compared to "ground truth." Ground truth is defined as the actual physical and cognitive workload experienced by the participant at any given moment. Ground truth is difficult to obtain in an experiment such as this one, for several reasons. Firstly, the experiment was conducted in a free play environment ñ scenarios were not scripted, tasks were not known *a priori*, and participants (and opposing players) were free to conduct the overall mission in whatever manner they chose, Thus, unlike a controlled experiment, the experimenters where not able to vary workload directly by imposing a rigid, well understood task structure. Secondly, performance measures where ill-defined at best, and did not offer a systematic method for deriving workload. Thirdly, there were limited opportunities to probe the participants during the missions to gauge their workload. Occasionally the company commander would (under experimenter direction), ask the PL or PSG for a ìsituation reportî (SITREP), and reaction time and content could be evaluated to gauge the participantís current workload. But this was not sufficient for the resolution of ground truth needed ñ a moment to moment assessment of the actual workload of the subject. Finally, the level of cognitive workload induced in a participant is a function of not only the task load, but of factors such as stress, fatigue, training, experience, and individual differences in capabilities. Thus there was no way to directly correlate task load to workload in a systematic way to derive ground truth.

Therefore a process was developed utilizing experts raters to review the data to make a subjective assessment of the participantís ground truth workload on a moment-to-moment basis. Multiple data sources were captured during the experiment. Raters reviewed video streams (continuous video of the subject and intermittent video of the platoon), taking into account the various other data sources (notes from central observer, real-time annotations from shadower, post-scenario cognitive walkthroughs with subjects, and questionnaires), to make a moment to moment assessment of the cognitive workload being experienced by the participant at any given timestamp. The result was a time stamped series of blocks of low, medium, or high cognitive workload. Physical load was also assessed by the experts.

For both physical and cognitive workload, states were labeled, and well as the expert assessment of workload.

For Physical load states included: crouching, prone, kneeling, seated, standing, milling, walking, running, lifting, climb up/down ladder, climb up/down stairs, dragging stuff/body. Physical load was assessed to be either HIGH (running, lifting, dragging) or LOW (everything else).

Cognitive states included planning, movement, giving/receiving orders, receiving information, clearing building, responding to enemy, respond to civilians, report, respond to action, defend, secure, request, maintain vigilance, prepare equipment, and after action review. Subtasks to these high level tasks were also identified. No attempt was made to be complete, nor was this list of states structured as taxonomy. Rather the states were derived bottomís up from the data sources, and the list served as a common vocabulary for expert raters. Cognitive workload was defined as HIGH or LOW. LOW cognitive workload was defined as participants doing very little, or mundane tasks, and they could easily take on additional tasks. HIGH cognitive workload was defined as the participant unable to take on any additional tasks or to handle current task load.

The expert raters considered the multiple data sources, and created a spreadsheet with moment-to-moment assessments of both cognitive and physical workload. The rules of the process dictated that a new row was created for the spreadsheet whenever the participant does any one of the following: charges physical state (e.g. transition from standing to crouching), starts a new task (e.g. starts responding to enemy while shouting orders), changes physical load (e.g. LOW to HIGH), or changes cognitive workload (e.g. HIGH to LOW).

Once two experts have reviewed the video, the individual ratings are combined into a spreadsheet and the degree of agreement is calculated on a second-by-second scale. Areas of disagreement are flagged for joint review. If the two raters cannot agree on a final workload coding, a third expert is asked to ìbreak the tieî. This option was never needed. The reconciled ratings are then finalized and used as the basis, or ground truth, of the classifier accuracy.

RESULTS

Ground Truth Inter-rater reliability

Two experts independently performed the ground truth analysis described above. For the data sets analyzed (about 9.8 hours of data), agreement between the raters was high (physical load 94.0%, cognitive workload 83.6%). A final, canonical, assessment of ground truth was created by reconciling of the two individual expert's assessments.

Classifier Accuracy

One strategy for dealing with momentary fluctuations in classification accuracy is to median filter the output of the classifier over different time windows. One consequence of temporal smoothing of classifier output is to introduce a lag in the decision process. Our analysis considered the trade off in accuracy with various temporal windows.

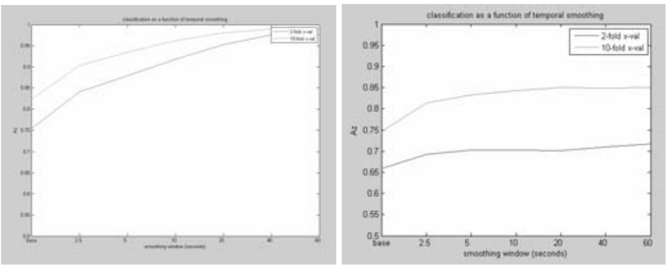

Figure 4. EEG based classification accuracy for the PL (left) and the PSG (right).

Figure 4 (left) illustrates base EEG classification accuracy for the platoon leader (PL) ranged from 76% (using 2-fold cross validation) to 83% (using 10-fold cross validation). Base results for the platoon sergeant (PSG) ranged from 66% (2-fold) to 75% (10-fold), as seen in Figure 4 (right). One strategy for robust classification in noisy field environments is to fuse data from multiple sources. Such an approach exploits the joint strengths of different data sources while minimizing their individual weaknesses. The fusion of cardiac (inter-beat interval) data provided a substantive boost to overall classification performance ñ these improvements were most pronounced for PSG, as seen in Figure 5. Base classification for PL went up to 87% (2-fold) and 95% (10-fold). Base classification for PSG went up from to 83% (2-fold) and 86% (10-fold) respectively.

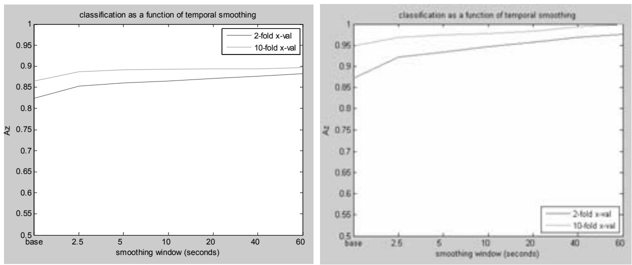

Figure 5. Classification accuracy for the fused sensor data for the PL (left) and the PSG (right).

DISCUSSION

With efficiency advances in signal processing and classification techniques, the paring down to the most effective and practical physiologically based sensing technologies, and the miniaturization of the sensing components, there has been a remarkable transformation from the laboratory based system to the current mobile classification ensemble. Developments in dry electrodes and helmet integration will further the capability to deploy these systems in operational environments. Additional work to further enhance the situational understanding of the individual soldiers will be to couple the cognitive state information with context aware sensors to truly gain the total picture. Context gathered from such sensors as accelerometers indicating body position and/or rifle position will be able to further inform whether the Soldier's current cognitive state is appropriate matched to the situation.

In conclusion, the authors believe this work represents the first demonstration of robust real-time cognitive state classification in the harsh operational MOUT environment. Furthermore, the workload classification accuracies obtained match that of the more pristine laboratory environment despite the motion, noise, and physical challenges posed by collecting physiological data in the field during real operations. Additionally, classification accuracy is equivalent to the inter-rater reliability between two expert human raters. The findings have implications for the use of physiological monitoring as a workload assessment tool to replace or enhance the use of more subjective tools such the NASA TLX and Cooper Harper ratings. This evaluation has proven that real-time workload assessment can be successfully used in the harsh and unforgiving military operational environment.

ACKNOWLEDGMENTS

This paper was supported by a contract with DARPA and the U.S. Army Natick Soldier Center, DAAD16-03-C-0054, for which Dr. Amy Kruse served as the Program Manager of the DARPA Improving Warfighter Information Intake Under Stress/AugCog program and Mr. Henry Girolamo was the DARPA Agent. The opinions expressed herein are those of the authors and do not necessarily reflect the views of DARPA or the U.S. Army Natick Soldier Center.

REFERENCES

Berka, C., Levendowski, C., Cvetinovic, M.M., Petrovic, M.M., Davis, G., Lumicao, M.N., Zivkovic, V.T., Popovic, M.V., & Olmstead, R. (2004). Real-time analysis of EEG indices of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human Computer Interaction, 17*(2), 151-170.

Duda, R.O, Hart, R.E., & Stork D.G. *Pattern Classification*, Second Edition. John Wiley & Sons, New York, 2001.

Garavan, H., Ross, T.J., Li, S.-J., & Stein, E.A. (2000). A parametric manipulation of central executive functioning using fMRI. *Cerebral Cortex, 10*, 585-592.

Gevins, A., & Smith, M. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex,* 10(9):829-39.

Makeig, S., & Jung, T-P. (1995). Changes in alertness are a principal component of variance in the EEG spectrum. *NeuroReport, 7*(1), 213-216.

Popivanov, D, & Mineva, A. (1999). Testing procedures for non-stationarity and non-linearity in physiological signals. *Mathematical Biosciences*, 157(1-2), 303-20.

Schmorrow, D.D., & Kruse, A.A. (2002). Improving human performance through advanced cognitive system technology. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSECí02)*, Orlando, FL.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature,* 381, 520-2.

Vapnik, V. (1999). The Nature of Statistical Learning Theory. Springer-Verlag.