RADIS: Real Time Affective State Detection and Induction System

Hesham Fouad
VR Sonic Inc.
Arlington, VA
hfouad@vrsonic.com

Ge Jin
Purdue University Calumet
Hammond, IN
ge.jin@purduecal.edu

ABSTRACT

Virtual Environment (VE) based immersive training systems have been widely adopted by US military as an alternative to costly and time consuming live training exercises. Current VE based training systems lack an affective state detection component, which may lead to decrements in training outcomes. In this paper, we introduce the Real-time Affective State Detection and Induction System (RADIS); a novel system that incorporates affective state detection and induction capabilities into existing training and simulation frameworks. RADIS is capable of: 1) dynamically monitoring a trainee's facial and speech features through visual and auditory channels, 2) detecting the trainee's affective state based on multimodal information fusion at the decision level, and 3) driving the trainee's affective state towards a target affective state specified in the structured lesson plan. RADIS currently uses human facial expression and speech sound for affective state detection. The visual and auditory signals are non-intrusive and provide higher prediction and recognition accuracy compared with physiological and motion signals. We extracted pitch, energy, formants, Mel-Frequency Cepstral Coefficients (MFCC), and speech rate from the speech signal and geometric and holistic features from the real time video input. The extracted feature vector was classified by the Support Vector Machine (SVM) to detect the trainee's affective state. RADIS was designed using a data-driven approach to support training domain independence. It follows the Sharable Content Object Reference Model (SCORM) eLearning standard and isolates all domain specific information in data so that a single code base can be successfully reused across multiple domains. By encoding all of the information required for a training session within the Structured Lesson Plan (SLP) the code base remains independent of the training domain and can be used in multiple training scenarios. RADIS will enhance the VE based training systems to better approximate the real-world experiences for the trainees.

ABOUT THE AUTHORS

Hesham Fouad is the President and Founder of VRSonic, Inc. He has over twenty five years of professional and academic experience in Computer Science. Dr. Fouad holds a B.S. in Audio Technology and M.S. in Computer Science from the American University and a D.Sc. in Computer Science from the George Washington University. His dissertation research was conducted at the Naval Research Laboratory and involved the development of perceptually optimal parallel real-time algorithms for synthetic media generation.

Ge Jin is currently an Assistant Professor in the Department of Computer Information Technology and Graphics at the Purdue University Calumet. He was a postdoctoral research scientist at the George Washington University department of computer science, from 2007 to 2008. His research spans the fields of computer graphics, virtual reality, computer animation, medical visualization, and educational game development. He is a member of the ACM SIGGRAPH, ASEE and International Society of Virtual Rehabilitation.

RADIS: Real Time Affective State Detection and Induction System

Hesham Fouad
VR Sonic Inc.
Arlington, VA
hfouad@vrsonic.com

Ge Jin
Purdue University Calumet
Hammond, IN
ge.jin@purduecal.edu

INTRODUCTION

Changes in modern warfare have resulted in frequent and lengthy deployments for US military personnel. These changes necessitate the development of training alternatives to the costly and time consuming live training exercises. One alternative to live training is the use of Virtual Environment (VE) based immersive training systems. Nevertheless, current VE based training systems focus exclusively on incorporating psychomotor and cognitive aspects of learning. This may lead to decrements in learning outcomes due to the lack of an affective learning component in those systems (Bloom et al., 1956). Recent findings from a wide spectrum of science including neuroscience, psychology, and cognitive science, indicate that emotion plays a critical role in a user's rational, functional, and intelligent behaviors (Picard, 1997). If training systems are to be effective, they must consider the user's affective state as a key component in learning. Through a process of affective state detection and induction, VE based training system can better approximate the real-world experiences encountered by soldiers upon deployment in the field. Integration of affective learning in a VE training system is an active area of research and a variety of approaches have been investigated. The related works include 1) the various sensory input channels that can be used to predict a user's affective state; 2) solution methods for extracting predictive features; 3) techniques for mapping those features to an emotional state; and 4) emotional state representation methods.

Different input channels have been used in human affective state detection research including visual, auditory, and physiological channels. The visual channel conveys facial expression, the auditory channel carries speech and vocal intonation and physiological channels carry various physiological signals. Most research efforts have focused primarily on the visual and auditory channels because of the intrusiveness and fragility of the physiological sensors and also because visual and auditory channels are better predictors of affective state when compared to physiological signals (Pantic & Rothkrantz, 2003). In the following sections we will discuss the relative merits and challenges in utilizing each of the above mentioned channels in affective state detection.

Human facial expression is known to be the most informative channel for human affective state detection (Sun et al., 2004). It has been reported that accuracy of affective state detection using facial expressions ranges from 70% to 98% (Pantic & Rothkrantz, 2003). Capturing human face image or image sequences is a non-intrusive process. The process of human affective state recognition from facial images involves translating a set of extracted facial features into a description of an affective state of the user. There are three steps involved in this process: 1) human face location tracking, 2) facial feature extraction, and 3) affective state classification. Approaches of facial feature extraction are categorized into three types. Facial Motion (Essa & Pentland, 1997; Otsuka & Ohya, 1998) is observed by using an optical flow method to describe the facial structure and a parametric representation of the face's muscle action groups is produced to be analyzed by a facial expression recognition process. Holistic Spatial Pattern (Edwards et al., 1998; Hong et al., 1998; Ji et al., 1999) detects facial expression based on the grey-level appearance of the face. This approach can produce good results for complicated images with variations in pose and lighting conditions. In Analytic Spatial Pattern detection (Colmenarez et al., 1999; Ebine et al., 2000; Pantic, 2001), the shapes of the various facial components such as the eyes, lips, brows, and lids as well as the relative distances between the components are analyzed in order to detect affective state. The main advantage of this approach is that it can be applied to partially occluded faces. All three extracted feature types can provide meaningful information for affective state detection. It is, therefore, possible to employ a combination of those features in order to improve recognition accuracy.

The auditory channel is considered to be a highly informative channel for human affective state detection. The non-verbal aspects of speech are conveyed in the manner in which it is spoken. This provides a rich source of information regarding human affective state and has therefore attracted a great deal

of attention by researchers. Recent efforts in affective state detection from speech signals have reported recognition accuracies ranging from 74% to 90% (Dellaert et al., 1996; Kang et al. 2000; Sato et al., 2001; Zhao et al., 2000). Affective state detection from speech signals involves the extraction of a set of features from the input speech signal followed by a classification stage where those features are mapped into a predicted affective state. The most widely used speech features include pitch, pitch contour, intensity, and speech rate (Pantic & Rothkrantz, 2003). Affective state detection from speech is a particularly attractive approach due to its relatively high predictive accuracy, low computational requirements, non-intrusive nature and low sensor cost. The primary drawback is that speech signals may be intermittent or not present at all based on the task being trained.

Physiological and autonomic nervous system activity in humans is closely related to affective state; particularly negative affective state (Rani and Sarkar, 2005). Various physiological measures are related to the level of physical or mental workload, vigilance, stress, or relaxation of the human being (Cai & Lin, 2007; Lin et al., 2007; Prendinger & Ishizuka, 2007; Wang et al., 2004). Physiological sensory measurements that have been used in human affective state detection research include: respiration rate, heart rate, heart rate variability, skin temperature, Galvanic Skin Response (GSR), skin conductance, electro-dermal activity, Electrocardiography (EEG), and Electromyography (EMG). The primary drawback of this approach is that it requires the use of intrusive and fragile sensors. This becomes particularly problematic for use in a deployed training system.

In practice, an operational human affective state detection system will most likely exhibit multimodality where different channels are used simultaneously and then combined to make a final assessment of affective state. By utilizing multiple channels, an affective state detection system can provide better performance, flexibility, and robustness (Wilamowitz-Moellendorff et al., 2005). Amongst the modalities commonly used for affective state detection, the visual and auditory channels have a number of advantages over physiological measures or motion; they provide the highest predictive capability for affective state detection and thus result in better recognition accuracy (Cowie et al., 2001). Additionally, these channels are more robust and generally utilize non-intrusive, low cost sensors. While advances in wearable computer technology have resulted in the development of less intrusive skin sensors for detecting physiological state (Mann, 1997), these sensors are fragile and sensitive to the amount of gel used in the application of the sensors making their use in an operational setting difficult (Cacioppo et al., 2000).

In this paper, we introduce the Real-time Affective state Detection and Induction System (RADIS) that, combined with Affective Virtual Environment Training System (A-VETS) design tool (developed by Design Interactive), provides an end-to-end scenario design and runtime training system that integrates affective learning into an immersive training system. The system is based on a Detection-Context-Induction model where the system dynamically monitors a trainee's affective state, compares that state to a target affective state encoded in a lesson plan for that training session, and then uses emotional state induction techniques to drive the trainees' emotional state towards a target state. Central to this model is a Structured Lesson Plan (SLP) specification. An SLP is a formal specification of training that acts as an interface between the A-VETS design tool and the RADIS runtime system. The SLP is generated by training instructors using A-VETS and then implemented at runtime by the RADIS system. RADIS can be applied to many types of virtual environment based training systems including war fighter training, medical crew training, and aviation training.

DESIGN OBJECTIVES OF RADIS

In this section, we outline a set of overarching design objectives for the RADIS system and the approaches that were used for achieving those objectives. These objectives guided the evolution of the RADIS system throughout this effort.

Training Domain Independence

The RADIS system was designed to be applicable to any training domain where immersive training is possible. This increases the utility of the system within DOD and enhances the potential for transition. Data driven systems isolate all domain specific information in data so that a single code base can be successfully reused across multiple domains. The gaming industry started moving towards this approach in order to enable the reuse of game engines across game titles. The SCORM eLearning standard, initially developed by the DOD, is another example of a data driven approach where a structured specification for content data is independent of the LMS implementation being used. By encoding all of the information required for a training session within the Structured Lesson Plan the code base remains independent of the training domain and can be used in multiple training scenarios. A Structured Lesson Plan is roughly equivalent to the Sharable Content Object specified in the SCORM 2004 standard. It is meant to encode, in XML data, all the elements necessary to specify both the static and

dynamic runtime behavior of the RADIS system during runtime.

The final SLP specification evolved through a continuous collaborative process with Design Interactive (DI). It consists of three major components: global session data, scenario selection rules, and scenario data. Global session data contains the information describing the whole training module, scenario selection rules determine the scenario sequence based on the given conditions, and scenario data contains the detailed information about each training scenario.

In order to achieve domain independence, the SLP has to encode domain expertise that enables the system to make reasonable decisions about how and when to modify the training scenario based on a trainee's current performance level and affective state. This domain expertise is encoded in the SLP using a set of rule-bases. Each training scenario contains two rule-bases that encode domain specific expertise for both individual and team training.

VE System Independence

The RADIS system interacts with an immersive simulation system in order to control scenario sequencing, dynamically inject media elements into scenes to induce affective state, and detect trainee actions in order to collect performance metrics. The current and future utility of the system will be severely limited if its design restricts its use to a single, currently available simulation system or standard. In order to address this problem, we defined a plug-in interface specification, as well as a software framework that enables the complete compartmentalization of simulation specific implementation details. The use of this approach enables the RADIS system to interact with nearly any simulation system or game engine.

Right Instruction at the Right Time

The development of a repository of training material that is accessible anywhere and that would provide individualized training when that training is needed is one of the goals of the development of the SCORM standard. A remotely accessible repository of SLPs can provide a distributed training capability that is accessible whenever needed. Another element however is needed to provide current user context (user details, current expertise, past performance on this training) so that the training is individualized. This context was kept locally or remotely in a repository.

AFFECTIVE STATE DETECTION

There are three types of input communicative channels that have been used in the human affective state detection research: visual, auditory, and physiological channels. The ideal solution for the human affective state detection will most likely exhibit multimodality in which numerous channels are used simultaneously and are combined to make a final assessment of affective state. The visual and auditory signals are non-intrusive and provide higher prediction and recognition accuracy compared with physiological and motion signals. Therefore, RADIS currently uses human facial expression and speech sound for affective state detection.

Emotional State Detection from Facial Features

In this effort, we used geometrical and holistic facial features for affective state detection. First, we detected the location and size of the frontal face from video input using a Haar cascade classifier. Inside the detected facial region, we localized left and right eyes, as well as mouth and nose positions. The contours of the eyes and mouth are extracted using corner detection and the convex hull method. In the geometric facial feature detection stage, we computed the eye and mouth open/closure states that are important for affective state detection. Second, the triangle defined by the center of mouth and eye positions is used to sample over 30 facial feature locations around the mouth, eyes, nose and forehead. Fifteen Gabor wavelet kernels (5 directions and 3 phases) are convoluted on the facial feature locations. The Gabor filter response at the geometric facial feature locations is used for pattern classifiers to detect the learner's emotional state.

The input images can vary in size. Our experiments indicate that the performance of the Haar object detection is dramatically affected by the size of image and minimum search kernel. A large image with a large face usually contains more facial information but also increases the detection time. Smaller faces usually have small eyes that are difficult to detect. To accurately detect the face in real time, we used a three level image pyramid. The original 640x480 pixel image was down-sampled to 320X240 and 160X120 images. The Haar face detection method was applied at the 160X120 size image. If the face detection is successful at lower resolution image, then we do not need to test on larger size images. This level-of-detail approach gave us real-time performance without degrading the face detection accuracy (Table 1).

After the face detection, we used a "face triangle" to describe a face for whole facial area analysis. The face triangle is an upside down triangle consisting of 3

points: left eyeball, right eyeball and mouth center. Once a face triangle is properly detected on a particular face, the exact face size can be estimated using that triangle. In addition, any potential feature point can be retrieved by conducting a linear calculation based on the triangle. A face triangle also can be evaluated to determine the correctness of the detection by measuring the ratio between base and the height of the triangle (BH ratio). Base is the distance between left and right eye, and height is the distance from mouth to the middle point between two eyes. The average BH ratio, based on our statistical analysis, is around 1.42 for 200 facial images examined. Any BH ratio smaller than 1.1 or larger 1.7 is considered a failed detection. In this case, our system fixes the face triangle by modifying the face ratio back to the normal range. The face triangle can also be used to determine the face tilting degree, which is important for localizing the derived feature locations from face triangle. In order to analyze the performance of our geometric feature localization method, we ran a number of facial image datasets through the system and collected the localization accuracy. We used a variety of publicly available datasets as well as a dataset created internally.

Table 1. Results of Geometric Feature Localization

Data	Num. Faces	Succeed	Fail	Accuracy
India	62	45	17	72%
JAFFE	224	220	4	98.2%
MMI	214	183	31	85.5%
VRSonic	85	71	14	83.5%
FEEDTUM	130	127	3	97.7%
Yale	164	123	41	75%
Total	879	769	110	87.4%

The results of our analysis are listed in Table 1. The geometric feature localization algorithm we developed has, on average, 87% localization accuracy. The two worst performing datasets were the Yale Face dataset and the India dataset. The Yale face dataset consists of facial images taken under extreme lighting conditions. It is expected then that the face localization accuracy is less than the average case. Based on our analysis of the India dataset, the average BH ratio for Indian people differs from the other datasets we analyzed. The BH ratio that we used as a normal ratio is 1.42. However, for Indian people, a BH ratio of 1.0~1.2 is common. In a deployed system, we must therefore adjust the BH ratio based on race. If we exclude the special case of the India and Yale Face datasets, the facial feature localization accuracy for our system becomes 92%. This level of accuracy was observed during our real-time webcam based affective state detection.

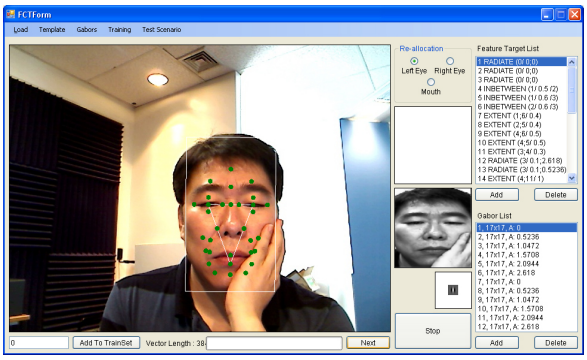

Figure 1. Face Template and Gabor Kernel Tool

Based on robust key facial feature locations, we derived the idea of a Facial Feature Template. A Facial Feature Template defines a set of critical positions relative to a "face triangle" that greatly affected by a facial expression change. We developed a face template editing and automatic feature vector generation tool so that we could experiment with classification parameter settings (Figure 1). With this tool, we can create derived facial feature locations, by clicking on the face image. The derived feature position is converted to a normalized relative location with regard to the three key feature locations forming the facial triangle. We used this tool to experiment with our method using different classification parameter settings. This enabled us to determine the parameters and methods that would result in optimal performance and detection accuracy.

In previous work done by (Bartlett et al. 2004), Gabor filters are convolved with an entire facial image in order to generate a feature vector. This dramatically increases the image processing and convolution time. Instead of convolving the entire face image, we applied a Gabor wavelet filter only to the Facial Feature Template locations described in previous section. The resultant feature vectors can be up to 105 in length. Although longer feature vectors do not affect the training time in SVM, which depends on the number of training datasets, we think shorter vectors that include only critical facial positions may win because they result in shorter prediction time and better pertinence. Furthermore, each key location involves only a one-time convolution (e.g. for a 17x17 filter, only 17x17 operations are carried out independent of image size). These two approaches significantly decrease the processing time and so we are able to maintain a detection rate of 24 frames per second on a standard PC. The Gabor wavelet convolution results generated

by this process are used as feature vectors for machine learning algorithms.

Based on the Face Feature Template and Gabor kernel convolution, we obtained a 384-element face feature vector. We analyzed the feature vector for each target emotional state using the Matlab software. When we plotted the face feature vector that corresponds to a Happy state for all the face datasets, we found that the facial state vector for that state was not easily distinguishable from other states and therefore would be difficult to detect. Instead of using the absolute values in the feature vectors, we used the difference between the each state vector and a face feature vector for a Neutral emotional state. This, in effect, generated a Differential State Vector that measures variance from a Neutral state. In the Differential State Vector for the Happy state, a number of feature elements became clearly distinguishable. This is due to the fact that, by using a differential vector we are analyzing the changes that occur in the face due to an emotional state and not the absolute state of the face. In order to integrate Differential State Vectors into our detection system, we introduced a 1-2 second calibration stage. During this stage, we instruct the trainee to maintain a Neutral emotional state while we collect 30-60 face images. These images are sent to a Neutral/Non-Neutral face classifier to determine if indeed we have a Neutral state. If we find that more than 40% of the images are Neutral faces, we use the mean feature vector of these Neutral faces as the base Neutral face vector. If we have less than 40% neutral faces, we revert to a pre-calculated mean Neutral feature vector as our base. During runtime, the difference between this base Neutral feature vector and the computed state vector is classified to detect Happy, Neutral and Angry affective states. Figure 2 shows the system in operation using the Differential Feature Vector.

Figure 2. Affective State Detection using Differential State Vector

We trained an SVM classifier using a number of facial image datasets including: the Japanese Female Facial Expression (JAFFE) Database (Lyons el al. 1998), the facial Expressions and Emotions from the Technical University Munich (FEEDTUM) Database (Wallhoff 2006) and MMI Facial Expression Database (Pantic et al. 2005). In addition, we also collected some facial expression images from the Internet. In scenario 1, we trained SVM classifier with 5 different affective states: Happy, Neutral, Angry, Sad, and Bored. We used 116 happy faces (52 Indian, 16 Japanese and 48 Caucasian faces), 64 neutral faces (16 Japanese and 48 Caucasian faces), 59 angry faces (31 Japanese and 28 Caucasian faces), 63 sad faces (31 Japanese and 32 Caucasian faces), and 106 bored images from Internet. Results show that our approach could accurately detect Neutral, Happy, and Angry affective states, with 90% correctness on both dataset and in a live camera test. All video clips have majority frames responding correctly, ranging from 50% to 70%. We developed a prototype detection system architected so that each module encapsulates one aspect of our research and we developed standard interfaces between them. By using and enhancing Haar object detection, eyeball tracking, face triangle analysis, and face template matching, we reached a high level of understanding for how such a process can be more efficient, accurate, and robust. Our results suggest that real-time classification of facial expression from images and video is achievable in real-time on standard PC hardware, and with good accuracy.

Affective State Detection from Speech

The second modality we used for affective state detection was human speech. The developed a system that uses a Support Vector Machine (SVM) algorithm trained using an emotional speech database. The system monitors a user's speech, analyzes the speech signal, and then determines their current affective state in real-time. It consists of three primarily components: signal preprocessing, feature extraction, and pattern classification. The pattern classification component uses the SVM algorithm, a data-driven approach where an SVM machine is trained using training data sets.

Preprocessing Dynamic Speech Detection

The speech preprocessing component is used to detect the onset of speech. It continuously monitors the user's speech signal via microphone and detects voice activity by determining if the current frame is speech or non-speech (noise). This is achieved by using a speech detection approach that dynamically adapts to the background noise level resulting in more robust performance than a traditional thresholding approach. (Ramírez et al. 2003, Cho and Kondoz 2001) The dynamic speech detection algorithm tracks the noise

characteristics and determines a speech threshold adaptively based on current noise characteristics. It computes the Speech threshold Ts from the estimated mean and variance of the log-energy of the noise.

$$T_s = \mu_n + \alpha\sigma_n \quad (1)$$

Feature Extraction

The feature extraction component of the speech detection system analyzes and extracts four different categories of speech features: pitch, MFCCs, formants, and energy. Those features are known to convey useful information about human affective state. Pitch information is obtained by applying pitch detection and pitch estimation algorithms. Each frame, this component first applies a low-pass filter to reduce the effects of the higher formants and any high frequency noise and applies a pitch detection algorithm based on autocorrelation. It finally applies a pitch estimation algorithm to remove the outliers and determine the pitch value for the frame. This feature extraction component of the system extracts ten features related with pitch information in total; they are mean, median, variance, maximum, minimum of the pitch values and their derivatives (Bänziger and Scherer 2005, Yu et al. 2001, Talkin 1995).

A formant is the spectral peak of the human voice signals meaning the acoustic resonance of the human vocal tract. The Linear Predictive Coding (LPC) method is used to model formants (Rabiner and Juang 1993, Petrushin 2000). The first three formants are estimated using LPC on 15 ms frames of speech. For each of the three formants, their derivatives and bandwidths, we calculate the mean, variance, maximum and minimum across all frames. We also calculate the mean, variance, maximum and minimum of the mean of each formant frequency, its derivative and bandwidth. The total number of formant related features are 48. The Mel-frequency Cepstrum coefficients (MFCCs) represent the short-term power spectrum of voice signals. Its frequency bands are equally spaced on the Mel scale so that it can approximate the human auditory system's response more closely than the linearly-spaced frequency bands used in normal Cepstrum. MFCCs are calculated by applying a short-time Fourier transform (STFT) algorithm on the input signal and transforming the powers of the spectrum into Mel frequency scale. Then the Discrete Cosine Transform (DCT) is applied on logs of the powers at each of the Mel frequencies (Kim et al. 2007). The first thirteen components of the amplitudes of the resulting spectrum are MFCCs. The system calculates mean, variance, maximum and minimum of all the frames of the first thirteen MFCCs. It also calculates the mean, variance, maximum and minimum of the mean of each coefficient and its derivative. The total number of MFCC related features are 112. The energy of the speech signals and speech rate are also used in the system. The energy of the signal is calculated by using Root Mean Square (RMS) method and speech rate is obtained by calculating the inverse of the average length of the speech part in each utterance.

Pattern Classification

The Support Vector Machine (SVM) model was developed and used as the pattern classifier for affective state detection from speech. The model was trained using the Emotional Prosody Speech database developed by University of Pennsylvania. The database consists of nine hours of speech data and contains speech in fifteen emotional categories. Among those fifteen emotional categories, the five affective states: Angry, Happy, Bored, Sad, and Neutral were used in training the SVM pattern classifier system. Training data sets were prepared separately for male and female voices respectively. The experimental result showed that the recognition accuracy was 87.42% for a male voice and 83.25% for a female voice. Table 3 and 4 show the recognition accuracy for each category of emotional state for male and female voice data.

Table 2. Recognition Accuracy for a Male and Female Voice

Affective States	Male	Female
Angry	91.93%	85.0%
Happy	90.31%	83.75%
Neutral	87.09%	92.5%
Bored	83.87	76.25%
Sad	83.87%	78.75%
Overall	87.42%	83.25%

AFFECTIVE STATE INDUCTION

The process of affective state induction involves the introduction of sensory cues into the simulated environment such that the trainee's affective state is driven to a target state specified by the instructional context. If we consider a set S consisting of all the sensory cues presented to a trainee at any time during a training exercise, then we can represent affective state induction as follows:

$$S_{total} = S_{training} + S_{affective} \quad (2)$$

Where S_{total} is the total set of auditory, visual and haptic cues presented to the trainee. $S_{training}$ is the set of cues related specifically to the training environment. In a MOUT (Military Operations on Urban Terrain) trainer, for example, these might include a visual

representation of the scene, the sound of weapons fire, and the visual and auditory representation of avatars. Finally, $S_{affective}$ consists of the set of cues introduced by the system solely to manage the trainee's affective state. The problem of affective state induction can then be stated as that of determining the contents of $S_{affective}$ either dynamically during a training exercise or statically during the design of the training scenarios. We can further distinguish the components of $S_{affective}$ if we consider some of the properties of human affect as well as the sensory cues that can induce affect.

Temporal Dimension of Affect

Human affective state can be segmented based on temporal properties of emotion into long term affect or mood and short term affect or emotion. A subject's mood is a persistent emotional state that exists over a long period of time (Ketal, 1975). In emotional space, mood can be viewed as the background onto which short term emotional responses are overlaid in the foreground. One of the best examples of the use of this distinction in managing affective state comes from the art of filmmaking. Filmmakers are expert in creating a general mood for a film. During a film, the induction of mood begins immediately at the start of the film and is driven by the careful selection of theme music, imagery and the visual graphics used for the title and credits. The result is that film viewers are immediately immersed into the mood of the story being told.

With respect to VE based training systems, the use of mood induction techniques can be considered as a static element of $S_{affective}$ that permeates throughout the training exercise. This induction can be achieved using techniques similar to those used in filmmaking including the introduction of musical elements in the environment, the selection of mood inducting colors in the scene visuals (e.g. bright red colors are known to induce stress) and finally audio and image manipulations can be introduced such as adding visual or auditory noise to the sensory cues being presented. Mood induction techniques are used currently in live training environments where trainers often play loud and often grating "Heavy Metal" music to induce stress in the trainees. If we integrate the temporal dimension into the definition of $S_{affective}$ we can now express that set of cues as follows:

$$S_{affective} = S_{mood} + S_{emotion} \quad (3)$$

Intrinsic Properties of Affective State Inducers

We now turn our attention to short term affective state inducers that can be used to manage a trainee's affective state during an exercise. We will limit our discussion to visual and auditory affective cues since very little is known about the impact of haptic cues on affective state. Affective state inducers can consist of either auditory or visual cues introduced during a training exercise. Examples of such cues are loud gunfire, colored noise, images of dead bodies or coloration of scene elements. We can see that some of the cues presented to a trainee are context dependent and are therefore specific to the training environment. For example, in a MOUT exercise the color of the interior walls of the simulated building can be manipulated to induce affective state. This approach would not be appropriate for a forward observer simulation since manipulating terrain color would not be practical. On the other hand, introducing colored noise into the simulated environment is context independent since it can be utilized in any training environment. We can therefore express $S_{emotion}$ as some combination of context dependent and context independent sensory cues as in equation 4, where D is the set of dependent cues and I is the set of independent cues.

$$S_{emotion} = D + I \quad (4)$$

In many regards, context independent cues may be advantageous because they can be easily introduced into any simulated environment without regard to context. Context dependent cues must be designed specifically for each training environment requiring additional effort. The one possible advantage of context dependent cues is that those cues may be relevant to the training task and therefore may have a positive impact on training transfer. The final property of affective state inducers that we consider is the continuity of the inducer. Affective state inducers may be inherently continuous so that the level of the cue can be manipulated in real-time. Examples of such cues are auditory noise, visual noise or musical elements. Each of those cues can be introduced into the environment in varying degrees to induce affective state. We refer to this class of cues as Continuous State Cues.

One important point regarding state continuous cues is that there is no known mapping between the amount of a continuous cue that is introduced and the corresponding shift in affect that results. It may be possible to utilize probabilistic models to "learn" such a mapping given that enough trails are used to train the model. The other class of cues are discrete cues which are either present or absent in the simulated environment. Such cues cannot be introduced in varying degrees into; they are either present or not. An example of a discrete cue is a visual element such as dead bodies that are placed in a scene. We refer to such cues as Discreet State Cues. Given the above distinctions, we can express the content of $S_{emotion}$ as in equation 5, where w is a scalar value indicating the

level of a continuous cue and o is a binary value indicating whether or not the cue is present.

$$S_{emotion} = wD_{\text{continuous}} + oD_{\text{discrete}} + wI_{\text{continuous}} \quad (5)$$

Affective State Induction in RADIS

Based on collaboration with researchers at Design Interactive, we developed a flexible and extensible approach for affective state induction in RADIS based on the idea of Emotion Induction Techniques (EIT). An EIT is a specification for a modification of the training environment that is known to drive trainees towards a target affective state. EITs can be context dependent (e.g. adding enemies to a room) or context independent (e.g. adding fog to the scene). Furthermore, EITs can be specified as individual training or team training EITs. EITs modify the training environment in real-time by invoking pre-designed scripts within the host simulation engine (Virtual Battle Space 2). These scripts were developed by Design Interactive as part of their AVETS tool and validated through human subject studies.

During runtime, the RADIS inference engine invokes EITs based on reasoning embedded within the current scenario's rule base. Execution of the inference engine is triggered each time one of the participants enters a previously unvisited scenario segment. Upon execution, a forward chaining reasoning process is carried out based on the current affective state and performance level of all the participants. The result of this execution is a set of candidate EITs. RADIS selects valid EITs for execution based on the following rules:

- EITs are never executed twice.
- An active EIT is never reactivated. If selected it is allowed to continue.
- Only one EIT is allowed to be active.
- Context dependent EITs are only activated in unoccupied scenario segments that have not been visited by any of the trainees. This avoids having training environment changes be detected by trainees.
- Team training EITs take precedence over individual training EITs

In order to support team training, the RADIS system evaluates two separate rule bases. A rule base for individual training drives an inferencing process on each trainee's machine and using the individual training rule base. This results in a set of EITs specific to that individual's affective state and performance. Concurrently to this, a rule base for team training drives an inferencing process on the RADIS IOS machine that generates EITs for each trainee based on team training intelligence. The system collects the team training EITs for each trainee and selects the set of EITs that have an occurrence rate amongst all trainees that exceeds a user specified minimum. If any team training EITs remain after this process, they are sent to each of the trainee's machines for implementation by the RADIS Engine.

RADIS SYSTEM

The RADIS system was developed as two components: the RADIS Engine and the RADIS Instructor Operator Station (RADIS IOS). The RADIS Engine contains functionality for affective state detection, performance metrics collection, rule-based reasoning, and communication with the IOS. The RADIS IOS component manages multiple trainees, displays trainee state, provides scenario execution control, performs rule-based reasoning for team training, and enables trainers to control system options. In the following sections we describe each of the RADIS components in more detail.

RADIS Engine

The RADIS Engine was developed as a plug-in component to the Virtual Battle Space (VBS2) serious game engine. We chose to use VBS2 as a target platform because it enjoys widespread use within the DOD training community and it is a mature software product. VBS2 also supports a software plug-in architecture that enables third-party vendors to extend its built-in capabilities with external modules. We utilized this approach to integrate affective learning within VBS2.

The RADIS Engine uses a callback mechanism as well as a script execution mechanism in order to communicate with VBS2. The callback mechanism invokes an entry point within the RADIS Engine each time a user enters or leaves a scenario segment (e.g. room, building, hallway). The script execution mechanism enables the RADIS Engine to initiate user scripts within VBS2 in order to implement EITs and to collect user performance metrics. Each time a user crosses a scenario segment boundary; the RADIS Engine performs the following series of operations in order to update its state and modify the training scenario:

1. Performance metrics are collected based on the specifications outlined in the Structured Lesson Plan (SLP) for that segment.
2. Trainee affective state is measured utilizing facial images, speech signals or both.

3. The inference engine is invoked using the individual training rule base specified in the SLP for that scenario segment. The execution of the inference engine results in a set of candidate EITs as well as potential modifications to state variables.
4. The resulting set of EITs is evaluated for possible execution based on the rules outlined in section 5.

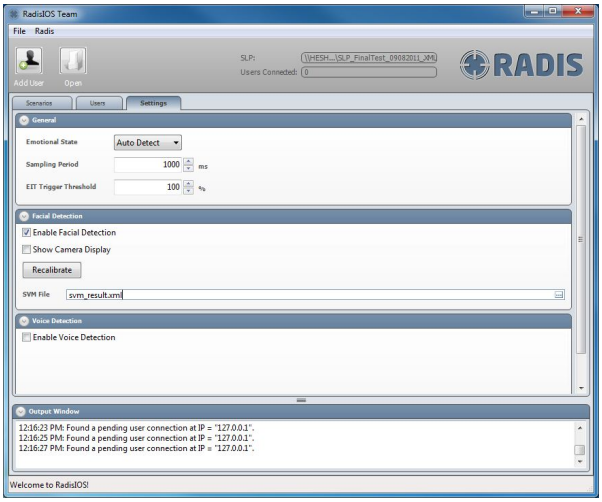

Figure 3. RADIS IOS Settings Panel

RADIS Instructor Operator Station

The RADIS IOS is a windowed application program that enables trainers to manage users, control training, view trainee status, and set system options. The application also performs the team training function in the RADIS System.

The main application area of RADIS IOS is a tabbed control consisting of three tabs: Scenario, Users, and Settings. It includes a settings tab (Figure 3) that gives the trainee fine grained control over the behavior of the RADIS system. The panel is divided into three topic areas: General, Facial Detection, and Voice Detection. The General section includes three settings. The Emotional State drop-down is used primarily for testing purposes. It enables the user to control how affective state detection is carried out. Users can let the RADIS Engine detect trainee emotional state automatically by selecting "Auto Detect" or they can select one of six states (Unknown, Happy, Angry, Sad, Bored, Neutral) as the current affective state for all users.

The next general option is the Sampling Period. This specifies the frequency with which RADIS IOS polls connected users to check on their state. This option is primarily used to avoid having RADIS IOS generate heavy network traffic by polling the remote computers too often.

The final general option is the EIT Trigger Threshold. As mentioned above, EITs generated by the team training rule base are first collected for each trainee, then based on their occurrence rate, are either selected as valid or discarded. The EIT Trigger Threshold determines the minimum occurrence frequency required for team training EIT to be considered valid. This parameter permits trainers to account for any outliers in a team of trainees and also to determine what constitutes team performance and affective state in a team training system.

CONCLUSION

In this paper, we presented a novel method to incorporate affective learning capability into virtual environment (VE) based training systems, in order to enhance existing training simulation frameworks. The proposed method is capable of monitoring a trainee's affective state and driving the trainee's emotional state towards a target emotional state specified in the lesson plan. Basic research was carried out to devise a structured approach for immersive training where affective learning could be an integral component. A multimodal affective state detection capability was devised that could operate using off-the-shelf web cameras and microphones making it usable in deployed settings. Finally, a complete immersive training system, RADIS, as developed and integrated into a commercially available serious game engine: VBS2. The technology is currently being customized for war fighter training simulators, it however can be applied to any other types of virtual environment based training systems such as medical crew training or aviation training.

The real-time detection of affective state is by no means a solved problem. Future work is required to improve the accuracy of the detection capability by improving the facial and voice detection capability and possibly by incorporating additional modalities such as body motion. This effort, in concert with Design Interactive's complementary efforts, has demonstrated that it is possible to integrate affective learning into current deployed training systems with plug-in interface.

ACKNOWLEDGEMENTS

This project was supported by Navy SBIR N0014-09-C-0063.

REFERENCES

Bänziger, T. and Scherer, K. R., (2005). The Role of Intonation in Emotional Expression, Speech Communication, Vol.46, 252-267.

Bartlett, M.S. Littlewort, G. Lainscek, C. Fasel, I. Movellan, J. (2004). Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions, University of California, San Diego, IEEE SMC, Hague, Netherlands, 2004.

Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956). Taxonomy of educational objectives: The lassification of educational goals. Handbook I: Cognitive domain. New York, Toronto: Longmans, Green.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., and Ito, T. A. (2000). The psychophysiology of emotion. Handbook of Emotions, Eds. New York: Guilford, pp. 173–191.

Cai, H. and Lin, Y. (2007). An experiment to non-intrusively collect driver physiological parameters towards cognitive/emotional state recognition. SAE 2007 World Congress, April 16-19.

Cho, Y.D. and Kondoz, A. (2001). Analysis and Improvement of a Statistical Model-based Voice Activity Detector, IEEE Signal Processing Letters, vol. 8, no. 10, pp. 276–278.

Colmenarez, A., Frey, B., and Huang, T. S. (1999). Embedded face and facial expression recognition. Proc. ICIP, 1999, vol. 1, pp. 633–637.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Mag., vol. 18, pp. 32–80.

Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. Proc. ICSLP, 1996, pp. 1970–1973.

Ebine, H., Shiga, Y., Ikeda, M., and Nakamura, O. (2000). The recognition of facial expressions with automatic detection of reference face. Proc. Canadian Conf. ECE, 2000, vol. 2, pp. 1091–1099.

Edwards, G. J., Cootes, T. F., and Taylor, C. J. (1998). Face recognition using active appearance models. Proc. ECCV, 1998, vol. 2, pp. 581–695.

Essa, I. and Pentland, A. (1997). Coding analysis interpretation recognition of facial expressions. IEEE Trans. Pattern Anal. Machine Intell., vol. 19, pp. 757–763, July

Hong, H., Neven, H., and Malsburg, C. (1998). Online facial expression recognition based on personalised galleries. Proc. FG, 1998, pp. 354–359.

Ji, S., Yoon, C., Park, J., and Park, M. (1999). Intelligent system for automatic adjustment of 3D facial shape model and face expression recognition. Proc. IFSC, 1999, vol. 3, pp. 1579–1584.

Kang, B. S., Han, C. H., Lee, S. T., Youn, D. H., and Lee, C. (2000). Speaker dependent emotion recognition using speech signals. Proc. ICSLP, 2000, pp. 383–386.

Ketal, R. (1975). Affect, mood, emotion, and feeling: semantic considerations American Journal of Psychiatry, 132, pp.1215-1217.

Kim, S., Georgiou, P., Lee, S., and Narayanan, S. (2007). Real-time Emotion Detection System Using Speech: Multi-modal Fusion of Different Timescale Features, Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece.

Lyons, M. J. Akamatsu, S. Kamachi, M. and Gyoba J., (1998). Coding Facial Expressions with Gabor Wavelets, Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, April 14-16 1998, Nara Japan, IEEE Computer Society, pp. 200-205.

Mann, S. (1997). Wearable computing: A first step toward personal imaging. Computer, vol. 30, no. 2, pp. 25–32.

Otsuka, T. and Ohya, J. (1998). Spotting segments displaying facial expression from image sequences using HMM. Proc. FG, 1998, pp. 442–447.

Pantic, M. (2001). Facial expression analysis by computational intelligence techniques. Ph.D. dissertation, Delft Univ. Technol., Delft, The Netherlands.

Pantic, M. Valstar, M.F. Rademaker R., and Maat, L., (2005). Web-based Database for Facial Expression Analysis, Proc. IEEE Int'l Conf. Multmedia and Expo (ICME'05), Amsterdam, The Netherlands, July 2005, DOI: 10.1109/ICME.2005.1521424

Pantic, M., & Rothkrantz, L.J.M. (2003). Toward and Affect-Sensitive Multimodal Human-Computer Interaction. Proceedings of the IEEE, 91(9), pp. 1370-1390.

Petrushin, V.A. (2000). Emotional Recognition in Speech Signal: Experimental Study, Development, and Application, ICSLP-2000, Vol.2, 222-225.

Picard. R. (1997). Affective Computing. Cambridge, MA: MIT Press.

Prendinger, H. and Ishizuka, M. (2007). Symmetric Multimodality Revisited: Unveiling Users' Physiological Activity. IEEE Transactions on Industrial Electronics, vol. 54, no. 2, pp. 692 – 698.

Rabiner, L.R. and Juang, B.H. (1993). Fundamentals of Speech Recognition, Upper Saddle River; NJ: Prentice-Hall, 1993.

Ramírez, J., Segura, J.C., Benítez, C., de la Torre, A., and Rubio, A. (2003). A New Adaptive Longterm Spectral Estimation Voice Activity Detector, Proc. EUROSPEECH 2003, Geneva, Switzerland, pp. 3041–3044.

Rani, P. and Sarkar, N. (2005). Operator Engagement Detection and Robot Behavior Adaptation in

Human-Robot Interaction. ICRA 2005, 18 – 22 April.

Sato, H., Mitsukura, Y., Fukumi, M., and Akamatsu, N. (2001). Emotional speech classification with prosodic parameters by using neural networks. Proc. Australian and NewZealand Intelligent Information Systems Conf., 2001, pp. 395–398.

Sun, Y., Sebe, N., Lew, M., and Gevers, T. (2004) Authentic emotion detection in realtime video. International Workshop on Human-Computer Interaction, Lecture Notes in Computer Science, vol. 3058, Springer (2004) 92–101.

Talkin, D. (1995). A Robust Algorithm for Pitch Tracking (RAPT), Speech Coding & Synthesis 1995.

Wallhoff, F., Facial Expressions and Emotion Database (2006)., Technische Universität München, 2006

Wang, H., Prendinger, H., and Igarashi, T. (2004). Communicating emotions in online chat using physiological sensors and animated text. CHI' 04, pp. 1171 – 1174.

Wilamowitz-Moellendorff, M., Muller, C., Jameson, A., Brandherm, B., and Schwartz, T. (2005). Recognition of time pressure via physiological sensors: Is the user's motion a help or a hindrance? Proceedings of the UM 2005 Workshop on Adapting to Affective Factors.

Yu, F., Chang, E., Xu, Y., and Shum, H. (2001). Emotion Detection from Speech to Enrich Multimedia Content, Lecture Notes In Computer Science, Vol.2195, 550-557.

Zhao, L., Lu, W., Jiang, Y., and Wu, Z. (2000). A study on emotional feature recognition in speech. Proc. ICSLP, 2000, pp. 961–964.