Analysis of Learners' State in an Integrated E-learning System

G. Caridakis, P. Tzouveli, A. Raouzaiou, K. Karpouzis and S. Kollias

Image, Video and Multimedia Systems Laboratory
9, Heroon Polytechneiou str
15780, Zografou, Athens, Greece
+30 210 7724053

{gcari, tpar, araouz}@image.ntua.gr, {kkarpou, stefanos}@cs.ntua.gr

ABSTRACT

A learning system adaptive to learner's behavior is considered as an innovative system. The more a learning system exchanges relevant fragments of information about the learner's affective status the more it adapts to it. Following this direction, we propose an integrated learning system taking under consideration learners' emotional state in order to provide a personalized e-learning system. An extended version of the IEEE Reference Model (WG) LTSA is used for this purpose. The proposed approach is based on the automatic analysis of the learners' emotional state providing different learners' profiles which are built and maintained by "observing" each learner behavior. As a learner is strongly positively affected to the learning procedure in the presence of an agent, the proposed system has adopted an expressive ECA (Embodied Conversational Agent) which is adapted to the learner's emotional states in the duration of the learning procedure.

Keywords: Decision Models, Knowledge Utilization, Information and Communication Technologies, Communications Software, Neural Networks, Human-Machine Systems, Electronic Learning (E-Learning), Motion Estimation, Multimedia Application

1. INTRODUCTION

Nowadays, it is widely accepted that information technology modifies the learning experience. Learning methods are becoming more and more portable, flexible, and adaptive. The WWW has been broadly adopted as a medium for network-enabled transfer of skills, information and knowledge and plays a significant role in all fields of education (Commission of European Communities, 2000). Web-oriented applications try to satisfy current educational needs, closing the gap between traditional educational techniques and future trends in technology-blended education. Towards this goal, various e-learning systems have been developed missing functionalities like educational multimedia environments, personalized capabilities and tracking of learners' input and relevance feedback (Karagiannidis, 2002). Though, tracking and grasping user behavior in real time remains the most challenging task to retrieve an appropriate and fine-grained user profile as well as to provide personalized learning content.

In this direction, we have adopted the IEEE Reference Model (WG) of the Learning Technology Standards Committee and especially an extension of it. In particular, this extension includes an analysis of learners' state as well as a profiling procedure. Moreover, concerning learner's emotional state, human computer interaction techniques are being employed. Sometimes, a simple facial expression or hand gesture, such as placing a person's hands over his ears, can pass on the message that he has had enough of what he is hearing; this is conveyed more expressively than with any other spoken phrase.

On one hand, application design must take into account the ability of humans to provide multimodal input to computers, thus moving away from the monolithic window-mouse-pointer interface paradigm and utilizing more intuitive concepts, closer to human niches (Jaimes, 2006; Petland ,2005). A large part of this naturalistic interaction concept is expressivity (Picard, 2000), both in terms of interpreting the reaction of the user to a particular event or taking into account their emotional state and adapting learning procedure to it, even for less technology-savvy users.

Since the early 1970s, Paul Ekman and his colleagues have performed extensive studies of human facial expressions. They found evidence to support universality in facial expressions. These "universal facial expressions" are those representing happiness, sadness, anger, fear, surprise, and disgust. They studied facial expressions in different cultures, including preliterate cultures, and found much commonality in the expression and recognition of emotions on the face. However, they observed differences in expressions as well, and proposed that facial expressions are governed by "display rules" in different social contexts. Additionally, the vocal aspect of a communicative message carries various kinds of information. If we disregard the manner in which the message was spoken and consider the verbal part (e.g., words) only, we might miss the important aspects of the pertinent utterance and we might even completely misunderstand what was the meaning of the message. Nevertheless, in contrast to spoken language processing, which has recently witnessed signifficant advances, the processing of emotional speech has not been widely explored. Finally, emotion consists of more than outward physical expression; it also consists of internal feelings and thoughts, as well as other internal processes of which the person having the emotion may not be aware. Still, these physiological processes can be naturally recognized by people. A stranger shaking your hand can feel its clamminess (related to skin conductivity); a friend leaning next to you may sense your heart pounding, etc.

On the other hand, the captivating presence of the agents can motivate learners to interact more frequently with agent-based educational software (Raouzaiou, 2002). To design the most effective agent-based learning environment, it is essential to understand how students perceive an animated pedagogical agent with regard to affective dimensions such as encouragement, utility, credibility, and clarity. In (Karpouzis et al, 2007), a study of the affective impact of animated pedagogical agents on learning experiences is presented. This study revealed the persona effect, which mentioned that the presence of a lifelike expressive character in an interactive learning environment can have a strong positive effect on student's perception of their learning experience. The study also demonstrates the interesting

effect of multiple types of explanatory behaviors on both affective perception and learning performance.

Consequently, the ability of virtual agents to provide expressive feedback to a learner is an important aspect to support their naturalness. In this framework, the proposed system provides learners' states analysis and an ECA-synthesis of multimodal cues constitutes an important part of an e-learning system.

The structure of the paper is as follows: Section 2 discusses related work, related both to ECAs and ECA enhanced integrated e-learning environments while Section 3 presents the overall system architecture as well as how the IEEE LTSA is extended in order to provide an adaptive learning system. Section 4, describes in detail the recording of the learners' state and more specifically the facial expression analysis, the hand detection and tracking and the gesture expressivity features extraction process. Finally, Section 5 concludes this work by summarizing the article, discussing the presented work and proposing future directions.

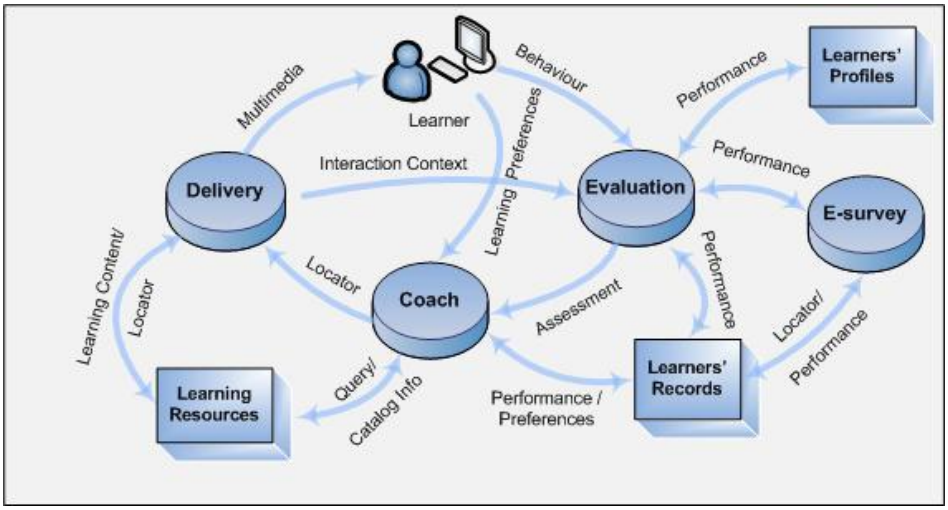


Figure 1: Extension of LTSA IEEE learning system

2. RELATED WORK

e-Learning, Affective Computing and Embodied Conversational Agents are extremely active research areas and in recent times several attempts have been made to integrate them in a unified context. These research areas benefit much more than merely discovering application areas in one another but new, extremely interesting problems, issues and points arise from such a constructive fusion.

2.1 ECAs

Embodied Conversational Agents (ECAs) are anthropomorphic virtual agents that have a human-like look and can interact with the user. Interactions can range from a textual input field and textual output for chatting with the agent to speech, gesture and even emotion-recognition and generation.

A student can learn interacting with one or more ECAs, who may provide information or encouragement, share knowledge, or collaborate with the student (Morishima, 2006; Baylor & Kim, 2005; Erickson, 1997). Agents can help overcome some constraints of conventional e- learning. Students exposed to an environment with an agent presented deeper learning and higher motivation than students without one (Mulken, 1998).

There are many examples of "teaching" agents: AutoTutor plays the role of tutor (Moreno, 2001), the agents Steve and Adele developed by CARTE are experts in naval training tasks (Graesser, 2001) while the agents developed by Baylor and Kim (Morishima, 2004) served distinct instructional purposes as an expert, a motivator, and a mentor. There is also the approach of *learning companion* (Johnson, 2000; Chan, 1990;0 Chou, 2003; Goodman, 1998; Hietala, 1998; Kapoor, 2005; Kim, 2003).

One of the issues in using ECAs as tutors is the ability to perceive and realistically express emotions during the interaction with the student. Social interaction is "the process by which we act and react to those around us" (Ryokai, 2003; Lang,1995) and consists of verbal and non-verbal communication. Non-verbal communication is performed via communication channels like body language, gesture and mimics (see Ryokai, 2003). As these communication channels all have to do with body movement the body has to look and behave human-like. Humans need believable and human-like behavior to trust such a virtual agent. For mimics good simulation of emotion is crucial to social interaction as showing emotions is part of our mimics (see Giddens, 2003; Ekman,1993).

Emotional models have increasingly become the center of interest in agent research over the last decade. This is attributable to the fact that one's own emotions play a crucial role in decision making (see e.g. Ekman, 1999), and that emotional information can easily be expressed by facial expressions. Moreover, it has become feasible to register the user's emotional state and integrate it in the agent's decision making process (see Bates 1994).

One approach that has proven successful and that is widely used for implementing an agent's emotional behavior is the OCC-model (Picard, 1997). A different approach that is widely employed makes use of a dimensional model of emotions (Ortony, 1988). Such a model is widely used for emotion recognition as well as emotion generation (Lang, 1995; Prendinger, 2004). Typically, emotions are defined by the two dimensions of arousal and valence.

Another important part for a credible virtual agent, especially in e-learning application, is how the virtual agents feel in a particular situation, how they react, while the transition between the different emotional situations (e.g. if they feel understood, accepted or not they get more friendly/ satisfied or angry) is also important.

Agents capable enhanced with the above mentioned features have successfully been tested in small-scale teaching scenarios. Agents in learning environments do not only allow for realistic communicative behaviors, but also for creating a sense of emotional involvement.

Pelachaud et al. (2002) focus on realistic facial and dialogue behaviors in an embodied conversational agent interacting with the user.

Hall and colleagues (2005) describe a virtual learning environment for children that makes use of autonomous agents. The agent's actions are triggered by an affective agent architecture that is used to appraise a situation and to suggest suitable ways of action depending on the agent's emotional state. By instructing an agent what to do next, the user is able to influence the enfolding story.

Hubal and colleagues (2000; 2003) developed a training software for law enforcement personnel that has to be able to recognize different behavioral signs (gestures, verbal cues) of mental illnesses like schizophrenia or paranoia.

Marsella and colleagues (2003) present an interactive pedagogical drama featuring a virtual character to teach problem solving skills to mother's of pediatric cancer patients. The agent is in the same situation as the mother and by interacting with the agent, the user is presented with some guidance on typical problems.

A number of different ways have been adopted by these systems to reconcile the ability of the agent to make choices about its next action and the requirement for a coherent narrative. In some cases, a branching approach is chosen, where user interaction follows one of a small number of finite paths. A more flexible approach adopted by Hall et al. (2005) is that of a Story Facilitator that shapes the story by selecting the initial conditions for episodes and using triggers to cause events within them (Aylett, 2006). An extension of this approach (Louchart, 2007) gives agents themselves the ability to evaluate the possible impact on others of their actions and select accordingly. Finally, the story-telling guide (Lim, 2007) constructs narrative from story-fragments and adds its own point of view as a commentary upon it.

Most of the above-mentioned systems can be classified as proof of concept, where a small part of a domain is modeled. All include some kind of emotional modeling - either for realistically displaying the agent or for affective reasoning processes and have shown the importance of including such a model in critical training applications. User interactions are mostly restrained to simple speech input and/or mouse interactions.

2.2 ECA enhanced integrated e-learning systems

Jung et al. (2005) developed an e-learning system enhanced with the presence of a virtual agent and focus more on the always sensitive situation of hospitalized children. Haptek virtual technology was employed and the affective virtual patient was multimodal in the sense that both highly emotional facial expression and lip-synchronized speech were controlled by the XML based modeling language used as the control system.

McQuiggan et al.(2008) study the possibilities of affective reasoning in a train/test scenario (CARE) where a trainer guides a virtual agent through a series of problem-solving tasks and then empathy models are used to drive runtime situation-appropriate empathetic behaviors by selecting suitable parallel or reactive empathetic expressions. Parallel empathy refers to mere replication of another's affective state, whereas reactive empathy exhibits greater cognitive awareness and may lead to incongruent emotional responses.

Virtual European School (VES) project (Bouras, 2001) also incorporates a communicative character of the multi-user distributed virtual environment that allows students and staff to meet in social shared spaces and engage in on-line real-time seminars and tutorials. This collaborative learning system that incorporates agents equipped with social intelligence is also presented by Morishima et al (2004). The effectiveness of the application was supported by a content quiz, according to which participants to the two co-learner conditions (Social Model and No Social Model) attained higher scores of correct answers than those in the no-agent condition. Interestingly enough the social model co-learner situation participants scored higher than the ones that participated a session where the agent had no social model associated with its behavior.

Fernandez-Caballero (2003) present Intelligent Tutoring System an application for enhancing e-learning / e-teaching. ITS consists of three components: the Student, Model and the Pedagogical Model. VBroker is presented by Laufer et al (2004) which is actually an E-learning curriculum and a virtual stock exchange game. Emphasis is given on the emotional model enabling the tutoring agent to provide proper emotional feedback and humorous behavior to achieve edutainment.

Finally, a collaborative affective virtual environment system is introduced by Neji & Ammar (2007). This technology offers the possibility of interlocutors to express themselves emotionally in an efficient and effective way, an animated virtual head, in the framework of EMASPEL (Emotional Multi-Agents System for Peer to peer E-Learning).

3. OVERALL ARCHITECTURE OF THE SYSTEM

In this section, the architecture of the proposed e-learning system is described. The proposed system includes three individual components which are considered as standalone entities, while, at the same time, collaborate with each other. The main entity, *Learner,* includes learners, participating in the learning process as individuals or in groups. The entity *Experts* includes experts in the e-learning field and computer engineers. The goal of that group is the design and continuous improvement of the system, with respect to computational integrity and educational efficiency. The Experts' Group is responsible for determining the learning topics that the e-learning schema can provide to the learners' group to which is addressed. The last entity is formed by the Server System, including all system hardware and software. Each subcomponent of this entity can be part of a centralized computer system or rely on separate hardware units scattered around the network. The internal architecture, in which the software of the proposed system has been based and developed, is considered as an extension of the IEEE Reference Model (WG) of the Learning Technology Standards Committee (LTSA). The IEEE LTSA (2001) contains three types of entities: processes (oval), process the information received from the stores entities via flows, stores (rectangles) implement an inactive system component used as an information repository and flows (vectors), denotes the transfer of information (control or data) from one entity to another. In the generic approach to e-learning systems outlined in the LTSA Draft Standard, system's ability to adapt its operation to the learner is not defined, although an evaluation process exists.

To handle this issue, we introduce in Figure 1 an extension of the IEEE LTSA enriching behavior flow (Mylonas et al, 2007) and all process modules with new functionalities, on the purpose of achieving effective collaboration between them. In the following, let us examine the role and the behavior of the different components in the proposed framework. Firstly, the group of experts defines the learning content in a specific learning subject which are accompanied with an affective Embodied

Conversational Agent (ECA). The learner content together with the ECA is illustrated to the learners through multimedia flow (Figure 1) on their first access to the system.

While a learning resource is presented to the learners, both behavior and reactions are observed. The behavior information, when a learner uses it for the first time, includes the recording of learner's reaction by a web camera which is placed in front of him. The learner entity's observable behavior and is given as input to the evaluation process which produces assessment information about the learner situation and sends it to the coach process. In addition, the evaluation process creates performance information flow stored in the learner records. Performance information can come from both the evaluation process (e.g. video recordings, learners' profile, answer to the questionnaire, grades on lessons) and the coach process (e.g. certifications). The learner records store hold information about the past (e.g. historical learner records), the present (e.g. current assessments for suspending and resuming sessions, current behavior) and the future (e.g. presentation rate, pedagogy, learner).

In order to extract learner profiles, the evaluation process takes into consideration the learner behavior and the learners' profile store. After that, the learner's level of knowledge in specific subject is computed, classifying the learner's profile, which is then stored in the learner records. In addition, we design a new store, i.e. learner profiles, which contain the current profiles of the e-learning system. Thus, once a learner is assigned to a learner profile, the coach uses information in order to locate the learning material from the learning resources store that best matches learner' profile and an appropriated agent is displayed to him according to his learning ability. Change of learners' profiles can be performed during their training, updating the learner records store.

The entity coach can receive performance information from the learner records at any time. Performance information, such as assessment information, certifications and preferences are stored in the learner records by the coach process. Based on this information, the coach requests the appropriate learning materials for each learner from the learning resources. Finally the delivery process transforms these materials via learning content store into a presentation which is accompanied with expressive ECA which is adapted to the learner emotional states.

4. RECORDING LEARNERS' STATE

A very important part of an educational system should be both recording and analysis of learner's state. The system becomes more efficient when it can be adjusted to every user. All learners don't concentrate for the same period of time or lose their interest when the rate of learning material is not very fast. On this section we analyze the way that the learner state is captured, intending to include this information to the learners' profile providing to the learner the proper learning material and in a proper learning rate.

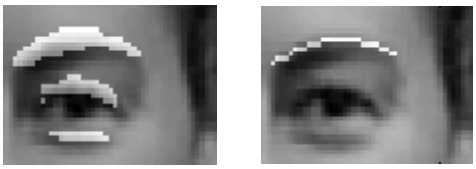
Active or passive, engaged or not, hyperactive or not active at all, positive or negative are some of possibly detected learner's states. Learner's state is extracted using face and hands gestures analysis together with body posture and motion. The detection of the learner's face and hands in the captured images is based on detection and evaluation of either skin segments or blobs that the captured images contain. For this purpose, the position and shape of mouth, eyes and eyelids are detected and features related to them are extracted.

Once the learner's state is recognized, the presentation of the learning content is tailored to the profile of each learner, while the system supports on-line adaptation to the learner's state through this multimodal analysis, being adapted to short term changes of the learner's state. Adaptation of the learner interface will be performed by getting relevant information from the current learner's profile and the current learner's state.

4.1 Facial Expression Analysis

Facial analysis includes a number of processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face (Karpouzis et al, 2007).

Facial analysis includes a number of processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face. Although FAPs provide all the necessary elements for MPEG-4 compatible animation, we cannot use them for the analysis of expressions from video scenes, due to the absence of a clear quantitative definition framework. In order to measure FAPs in real image sequences, we have to define a mapping between them and the movement of specific facial definition parameters (FDPs) feature points (FPs), which correspond to salient points on the human face.

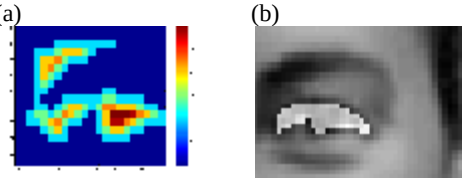


(a) (b)

Figure 2 (a) eyebrow-candidates. (b) selected eyebrow mask

The face is first located, so that approximate facial feature locations can be estimated from the head position and rotation. Face roll rotation is estimated and corrected and the head is segmented focusing on the following facial areas: left eye/eyebrow, right eye/eyebrow, nose and mouth. Each of those areas, called feature-candidate areas, contains the features whose boundaries need to be extracted for our purposes. Inside the corresponding feature-candidate areas precise feature extraction is performed for each facial feature, i.e. eyes (Figure 3), eyebrows (Figure 2), mouth and nose, using a multi-cue approach, generating a small number of intermediate feature masks. Feature masks generated for each facial feature are fused together to produce the final mask for that feature. The mask fusion process uses anthropometric criteria to perform validation and weight assignment on each intermediate mask; each feature's weighted masks are then fused to produce a final mask along with confidence level estimation. The edges of the final masks are considered to be the extracted feature points as can be seen in Figure 2a.

(a) (b)



(c) (d)

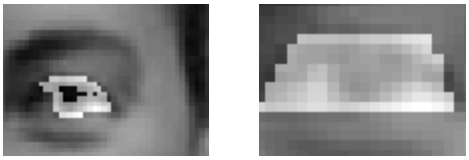


Figure 3: Eye masks

Measurement of FAPs requires the availability of a frame where the subject's expression is found to be neutral. This frame will be called the neutral frame and is manually selected from video sequences to be analyzed or interactively provided to the system when initially brought into a specific user's ownership. The final feature masks are used to extract 19 Feature Points (FPs); Feature Points obtained from each frame are compared to FPs obtained from the neutral frame to estimate facial deformations and produce the Facial Animation Parameters (FAPs). Confidence levels on FAP estimation are derived from the equivalent feature point confidence levels. The FAPs are used along with their confidence levels to provide the facial expression estimation.

Concerning the robustness of the proposed algorithm we had a subset of the available dataset manually annotated by a number of reviewers and compared this annotation versus the automatic one obtained by the described system. As a metric for robustness evaluation we incorporated a modified Williams' Index where the results of automatic feature extraction are considered as one of the reviewers. When WI is larger than 1, the computer generated mask disagrees less with the observers than the observers disagree with each other. Figure 4 and Figure 5 show the distribution of WI for eyes/mouth and eyebrows regions respectively. These results indicate that the described algorithm is efficient given the image is of acceptable quality, the head pose is quite frontal so that feature occlusion does not occur.

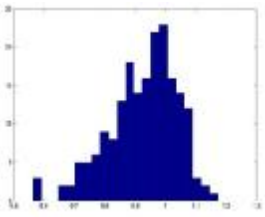


Figure 4: Williams Index distribution

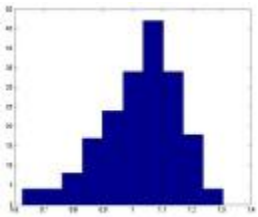


Figure 5: Williams Index distribution

Figure 6: Facial feature points on one of the subjects

In terms of affective state recognition a dynamic approach was adopted (Cowie et al, 2008), where a recurrent neural network, whose short-term memory and approximation capabilities cater for modeling dynamic events and classifying input patterns to emotional states. In order to consider the dynamics of displayed expressions a classification model that is able to model and learn dynamics is employed. Whereas conventional neural networks deal with static patterns the first layer of an Elman network has a recurrent connection adding a delay which stores values from the previous time step which can be used in the current time step, thus providing the element of memory. Finally the combination of facial expressions with other modalities would allows us to capture the users' emotional state, relying on the best performing modality in cases where one modality suffers from noise or bad sensing conditions (Castellano et al., 2007).

4.2 Head and Hand tracking

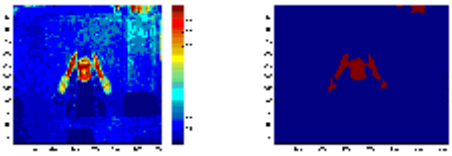
Several approaches have been reviewed for the head-hand tracking module. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin detection and tracking module. The general process involves the creation of moving skin masks, namely skin color areas that are tracked between subsequent frames. By tracking the centroid of those skin masks, we produce an estimate of the learner's movements. A priori knowledge concerning the human body and the circumstances when filming the gestures was incorporated into the module indicating the different body parts (head, right hand, left hand).Subsequently we expect to locate the head in the middle area of the upper half of the frame and the hand segments near the respective lower corners. The upper and lower segments of the hands are not detected or tracked.

The general process involves the creation of moving skin masks, namely skin color areas that are tracked between subsequent frames. By tracking the centroid of those skin masks, we produce an estimate of the user's movements. A priori knowledge concerning the human body and the circumstances when filming the gestures was incorporated into the module indicating the different body parts (head, right hand, left hand). For each frame a skin color probability matrix is computed by calculating the joint probability of the Cr/Cb image values. The skin color mask is then obtained from the skin probability matrix using thresholding. Possible moving areas are found by thresholding the pixels' difference between the current frame and the next, resulting in the possible-motion mask. This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color and motion masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. All described morphological operations are carried out with a disk-structuring element with a radius of 1% of the image width. The distance transform of the color mask is first calculated and only objects above the desired size are retained. These objects are used as markers for the morphological reconstruction of the initial color mask. The color mask is then closed to provide better centroid calculation. For the next frame, a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area. In the case of hand object merging and splitting, e.g., in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand. The Sagittal plane information of the gesture was ignored since it would require depth informa-

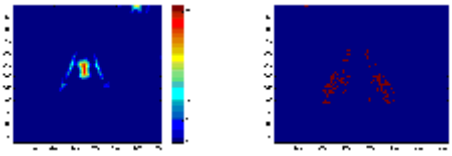
tion from the video stream and it would make the performance of the proposed algorithm very poor or would require a side camera and parallel processing of the two streams. The described algorithm is lightweight, allowing a rate of around 12 fps on a usual PC during our experiments, which is enough for continuous gesture tracking. The object correspondence heuristic makes it possible to individually track the hand segments correctly, at least during usual meaningful gesture sequences. In addition, the fusion of color and motion information eliminates any background noise or artifacts, thus reinforcing the robustness of the proposed approach.

The tracking algorithm is responsible for classifying the skin regions in the image sequence of the examined gesture based on the skin regions extracted from the described method. Skin region size, distance with reference to the previous classified position of the region, flow alignment and spatial constraints. These criteria ensure that the next region selected to replace the current one is approximately the same size, close to the last position and moves along the same direction as the previous one as long as the instantaneous speed is above a certain threshold. As a result each candidate region is being awarded a bonus for satisfying these criteria or is being penalized for failing to comply with the restrictions applied. The winner region is appointed as the reference region for the next frame.

Original Frame



(a) (b)



(c) (d)

Figure 7: Key steps in hand detection and tracking (a) Skin probability (b) Thresholding & Morphology operators (c) Distance transformation (d) Frame Difference

The criteria don't have an eliminating effect, meaning that if a region fails to satisfy one of them is not being excluded from the process, and the bonus or penalty given to the region is relative to the score achieved in every criterion test. The finally selected region's score is thresholded so that poor scoring winning regions are excluded. In this case the position of the body part is unchanged with reference to that in the previous frame. This feature is especially useful in occlusion cases when the position of the body part remains the same as just before occlusion occurs. After a certain number of frames the whole process is reinitialized so that a possible misclassification is not propagated. Head and head occlusion is tackled with the following simplistic yet efficient method. Suppose occlusion occurs at frame n and ceases to exist at frame k. The position of the hand during the occlusion phase (frames n-k) is considered to be the position of the hand at frame n-1. After frame k the detection and tracking algorithm for the specific hand continues normally.

4.3 Gesture Expressivity Features Extraction

Expressivity of behavior is an integral part of the communication process as it can provide information on the current emotional state, mood, and personality of a person (Wallbott & Scherer, 1986). Many researchers have investigated human motion characteristics and encoded them into dual categories such as slow/fast, small/expansive, weak/energetic, small/large, unpleasant/pleasant. To model expressivity, in our work, we use the six dimensions of behavior described by Hartmann et al., (2005), as a more accomplished way to describe the expressivity, since it tackles all the parameters of expression of emotion. Five parameters modeling behavior expressivity have been defined at the analysis level, as a subset of the above-mentioned six dimensions of behavior (see also next Section):

- Overall activation
- Spatial extent
- Temporal
- Fluidity
- Power

Overall activation is considered as the quantity of movement during a conversational turn. In our case it is computed as the sum of the motion vectors' norm: $A = \sum_{i=1}^{N} |r|^2 + |l|^2$. Spatial extent is modeled by expanding or condensing the entire space in front of the agent that is used for gesturing and is calculated as the maximum Euclidean distance of the position of the two hands: . The average spatial extent is also calculated for normalization reasons. The temporal expressivity parameter of the gesture signifies the duration of the movement while the speed expressivity parameter refers to the arm movement during the gesture's stroke phase (e.g., quick versus sustained actions). Gestures have three phases: preparation, stroke and retraction. The real message is in the stroke, whilst the preparation and retraction elements consist of moving the arms to and from the rest position, to and from the start and end of the stroke. Fluidity differentiates smooth/graceful from sudden/jerky ones. This concept seeks to capture the continuity between movements, as such, it seems appropriate to modify the continuity of the arms' trajectory paths as well as the acceleration and deceleration of the limbs. To extract this feature from the input image sequences we calculate the sum of the variance of the norms of the motion vectors. Power actually is identical with the first derivative of the motion vectors calculated in the first steps.

4.4 Expressive Synthesis

Communicative capabilities of conversational agents could be significantly improved if they could also convey the expressive component of physical behavior. Starting from the results reported in [Wallbott, 1986], we have defined and implemented [Hartmann, 2005a] a set of five parameters that affect the quality of the agent's behavior, that is the movement's spatial volume

(SPC), speed (TMP), energy (PWR), fluidity (FLT) and repetitivity (REP). Thus, the same gestures or facial expressions are performed by the agent in a qualitatively different way depending on this set of parameters. The Spatial Extent (SPC) parameter modulates the amplitude of the movement of arms, wrists (involved in the animation of a gesture), head and eyebrows (involved in the animation of a facial expression); it influences how wide or narrow their displacement will be during the final animation. For example let us consider the eyebrows raising in the expression of surprise: if the value of the Spatial Extent parameter is very high the final position of the eyebrows will be very high in the forehead (i.e. the eyebrows move under a strong of muscular contraction). The Temporal Extent (TMP) parameter shortens or lengthens the motion of the preparation and retraction phases of the gesture as well as the onset and offset duration for facial expression. On of the effect on the face is to speed up or slow down the rising/lowering of the eyebrows. The agent animation is generated by defining some key frames and computing the interpolation curves passing through these frames. The Fluidity (FLT) and Power (PWR) parameters act on the interpolation curves. Fluidity increases/reduces the continuity of the curves allowing the system to generate more/less smooth animations. Let us consider its effect on the head: if the value of the Fluidity parameter is very low the resulting curve of the head movement will appear as generated through linear interpolation. Thus, during its final animation the head will have a jerky movement. Power introduces a gesture/expression overshooting, that is a little lapse of time in which the body part involved by the gesture reaches a point in space further than the final one. For example the frown displayed in the expression of anger will be stronger for a short period of time, and then the eyebrows will reach the final position. The last parameter, Repetition (REP), exerts an influence on gestures and head movements. It increases the number of stroke of gestures to obtain repetition of the gestures themselves in the final animation. Let us consider the gesture "wrists going up and down in front of the body with open hands and palms up", a high value of the Repetition parameter will increase the number of the up and down movements. On the other hand this parameter decreases the time period of head nods and head shakes to obtain more nods and shakes in the same lapse of time.

5. Architecture evaluation

A formative evaluation of the so far implemented system is judged of crucial importance in order to decipher how the synthesized behavior is perceived and thus identify ways to improve the current output produced. Feedback on expressivity makes sense both within and out of context [Hartmann et al., 2005b]. At this point we are going to present a scheme on ways we propose in order to evaluate current work. Due to strict time limitations the results of this ongoing evaluation scheme will be made available in future publications.

5.1 Evaluation scheme

There are various questions worth answering, through rating tests, regarding the perception of synthesized emotional expressions in order to continue an in depth analysis of the parameters just mentioned. In our case, a first question is the perception and classification of the synthesized output by human viewers. We plan to conduct a rating test where twenty postgraduate students will be asked to rate a sequence of videos – one at a time. They will be presented a questionnaire comprised of scale questions regarding the percent of each class they believe the video snippet they just watched belongs to. The choice of classes is dictated by the natural language scenarios given to the actors of the original videos.

Participants will be asked how synthesised animations are perceived both in and out of the originating actor context by viewing the videos of the behavior mimicry scenarios. Thus, they will be divided into two groups, and will be randomly assigned to two conditions; the first group will be presented with the synthesised videos whereas the second group will be presented with both acted videos and their synthesised equivalents. The playback order of each set of videos will be chosen randomly so as to avoid ordering effects. In each group we plan to include a neutral condition of Greta simply blinking and looking straight ahead with a neutral expression as our conceptual baseline.

Questionnaires will also be used to collect participant feedback on which emotion they identified in each of the sequences they viewed. Their replies will be constrained with labels. Participants' confidence shall again be measured indirectly by asking them questions about the intensity of the perceived emotion.

Results from this first rating test can provide useful data on the perception and recognition of the synthesised expressions, as well as information on the effect of context (acted video sequences) in affective state perception of synthesized mimicked versions. Confidence measures will help draw conclusions on the role of the expressivity parameters and further refined manipulation of these parameters in conjunction with new rating tests can help decipher the role of these parameters in the perception of synthesized expressions.

5.2 Preliminary results

Although a formal evaluation of the overall architecture is not yet implemented, evaluating intermediate components has been carried out. We investigated the correlation of the automatic analysis results with the participants' results for each expressivity parameter (Table 1). Fist we took into account all 16 sets of ratings for each participant. Then we also calculated correlation coefficients when only taking into account ratings for the last 12 videos that each user viewed based on the idea that users had a by then developed a clearer idea of what they were doing. In both cases we correlated both the average of users' scores for each expressivity parameter of each video (rows marked as 'Avg' in Table 1 for 16 and 12 videos respectively), as well as their median scores (rows marked as 'Median' in Table 1 for 16 and 12 videos respectively).

	Ov.Act.	**Sp.Ext**	**Speed**	**Fluid**	**Power**
Avg/16	0.38	0.78	0.53	-0.20	0.57
Median/16	0.38	0.76	0.50	-0.21	0.61
Avg/12	0.36	0.74	0.50	-0.22	0.52
Median/12	0.38	076	0.48	-0.18	0.55

Table 1 Correlation coefficients of users' expressivity parameters against machine extracted values.

From these results it is apparent that there is no significant variance between the different data views investigated. We measured significant correlation only in the case of the special extent feature. Power and speed also correlated relatively well. The low values for overall activation and the negative correlation for the case of fluidity show that user's did not conceive these parameters correctly. These values might also be caused by the fact that the videos were of particularly short duration.

6. CONCLUSIONS

Present work introduces an affective approach to an e-learning system adaptive to the learner's emotional state bringing together established and emerging research areas such as e-Learning, Affective Computing and Embodied Conversational Agents. Extending the IEEE Reference Model (WG) LTSA, by enriching behavior flow and modules with new functionalities,

with the purpose of achieving effective collaboration between them, the proposed system is enriched with emotional awareness and by taking under consideration the learners' emotional state provides a personalized e-learning system both in terms of content and presentation. Affective awareness is achieved via facial expression recognition and extraction of gesture expressivity features while the ECA, shouldering the role of the teacher, attempts to adapt its behavior accordingly. The persona effect is then maximized since the learning procedure in the presence of an agent especially an expressive and life-like one. In this framework, the proposed system provides learners' states analysis and an ECA-synthesis of multimodal cues constitutes an important part of an e-learning system.

Future work includes the integration of a pedagogical module which will cater the reasoning of the ECA's behavior according to the cues received from the user and processed by the proposed system. Such a process will ensure that the ECA's behavior complies with widely accepted pedagogical methods and goes beyond an expressive agent.

An evaluation study is also essential for this kind of architectures and part of our future plans for the proposed system. User feedback and systematic monitoring of the students' curriculum performance will be collected through questionnaires and student grades respectively for different tutoring scenarios. The implemented scenarios will include traditional e-learning tutoring system, ECA enhanced tutoring system and finally the affective ECA tutoring system. Thus the system's efficiency will be evaluated and appropriate redesign adjustments will be carried out.

7. REFERENCES

Aylett, R.S, Figueiredo, R, Louchart, S, Dias, J. & Paiva, A. (2006). Making it up as you go along – improvising stories for pedagogical purposes. In: Gratch, J, Young, M, Aylett, R, Ballin, D. & Olivier, P (Eds) *6th International Conference, IVA*, (pp. 307-315) Springer, LNAI 4133.

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 15(2) 95-115.

Bouras, C., Philopoulos, A. & Tsiatsos, T. (2001). e-Learning through distributed virtual environments, *Journal of Network and Computer Applications* 24(3), 175-199.

Castellano, G., Kessous, L. & Caridakis, G., (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. In Peter, C., Beale R. (Eds), *Affect and Emotion in Human-Computer Interaction*, Lecture Notes in Computer Science, Springer

Cowie, R., Douglas-Cowie, E., Karpouzis, K., Caridakis, G., Wallace, M. & Kollias, S. (2008). Recognition of Emotional States in Natural Human-Computer Interaction. In Tzovaras D. (Ed) *Multimodal User Interfaces*, (pp. 119-153), Springer Berlin Heidelberg.

Chan, T. W., & Baskin, A. B. (1990). Learning companion systems. In C. Frasson & G. Gauthier (Eds.), *Intelligent tutoring systems at the crossroads of artificial intelligence and education.* (pp. 7–33): NJ: Ablex Publishing Corporation

Chou, C. Y., Chan, T. W., & Lin, C. J. (2003). Redefining the learning companion: The past, present, and future of educational agents. *Computers & Education*, 40, 255–269.

Commission of European Communities, *Communication from the Commission: e-Learning – Designing Tomorrow's Education*, Brussels, 2000

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.*

Ekman, P. (1999). Basic Emotions. In T. Dalgleish & T. Power (Eds): *The Handbook of Cognition and Emotion*. (pp. 45-60), John Wiley and Sons, 1999.

Ekman, P.: Facial Expressions and Emotion. *American Psychologist*, 48(4), 384-392, 1993.

Erickson, T. (1997). Designing agents as if people mattered. In J. M. Bradshaw (Ed.), *Software agents* (pp. 79–96). Menlo Park, CA: MIT Press.

Fernandez-Caballero, A., Lopez-Jaquero, V., Montero, F. & Gonzalez, P., (2003). Adaptive Interaction Multi-agent Systems in E-learning/E-teaching on the Web Lecture Notes In *Computer Science*, (pp.144-153), Springer.

Giddens, A., Duneier, M., & Appelbaum, R. P. (2003) *Introduction to Sociology*. 4th edition.

Goodman, B., Soller, A., Linton, F., & Gaimari, R. (1998). *Encouraging student reflection and articulartion using a learning companion.* Paper presented at the 8th International Conference on Artificial Intelligence in Education, Kobe, Japan.

Graesser, A. C., Person, N. K., Harter, D., & Group, T. R. (2001). Teaching tactics and dialog in autotutor. *International Journal of Artificial Intelligence in Education*, 12, 257– 279

Hall, L., Woods, S., Aylett, R. Newall, L. & Paiva, A. (2005). Achieving empathic engagement through affective interaction with synthetic characters. *Affective computing and intelligent interaction,* Springer.

Hartmann, B., Mancini, M., Buisine, S., and Pelachaud, C. (2005b). *Design and evaluation of expressive gesture synthesis for embodied conversational agents*. AAMAS'05. Utretch.

Hietala, P., & Niemirepo, T. (1998). The competence of learning companion agents. *International Journal of Artificial Intelligence in Education*, 9, 178–192.

Hubal, R. C., Frank, G. A., & Guin, C. (2003). Lessons Learned in Modeling Schizophrenic and Depressed Responsive Virtual Humans for Training. *Proceedings of International Conference on Intelligent User Interfaces* (pp. 85 – 92), ACM

Hubal, R. C., Kizakevich, P. N., Guinn, C. I., Merino, K. D., & West, S. L.:(2000). The Virtual Standardized Patient – Simulated Patient-Practitioner Dialog for Patient Interview Training. In *Envisioning Healing: Interactive Technology and the Patient-Practitioner Dialogue*, (pp.133-138) IOS Press: Amsterdam.

Jaimes A. (2006). Human-Centered Multimedia: Culture, Deployment, and Access, *IEEE Multimedia Magazine*, Vol. 13, No.1.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Faceto- face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11, 47–78.

Bates, J. (1994). *The role of emotion in believable agents* Communications of the ACM, 37(7):122–125, 1994.

Jung, B., Ahad, A. & Weber, M. (2005). The Affective Virtual Patient: An E-Learning Tool for Social Interaction Training within the Medical Field *Proceedings of International Conference TESI-Training Education & Education.*

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. *Proceedings of the 13th annual ACM international conference on Multimedia*, (pp.677-682), ACM New York, NY, USA

Karagiannidis, C., Sampson, D.G., & Cardinali, F.(2002). An architecture for Web-based e-Learning promoting reusable adaptive educational e-content, *Educational Technology & Society*, 5(4).

Karpouzis K., Caridakis G., Kessous L., Amir N., Raouzaiou A., Malatesta L., Kollias S., (2006). Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In T. Huang, A. Nijholt, M. Pantic, A. Pentland (eds.), *Lecture Notes in Artificial Intelligence, Special Volume on AI for Human Computing*, pp. 91-112, Springer.

Kim, J., André, E., Rehm, M., Vogt, T., & Wagner, J. (2005). Integrating Information from Speech and Physiological Signals to Achieve Emotional Sensitivity, in: *Proceedings of Interspeech/Eurospeech*, 2005.

Kim, Y. (2003). *An agent as a learning companion: What it matters*. Paper presented at the Annual Conference of Association for Educational Communications and Technology, Anaheim, CA.

Lang, P. J. (1995). *The Emotion Probe: Studies of Motivation and Attention*, American Psychologist 50 (5) 372–385.

Laufer, L. & Tatai, G. (2004). Learn, Chat & Play-An ECA Supported Stock Markets E-Learning Curricula'. *Proceedings of the IASTED International Conference on Web-Based Education* (pp.16-18) ACTA Press.

Lester J.C., Converse S.A., Kahler S.E., Barlow S.T.,. Stone B.A & Bhogal R.S. (1997). The persona effect: affective impact of animated pedagogical agents, *Proceedings of the SIGCHI conference on Human factors in computing systems*, (pp.359-366), ACM New York, NY, USA

Lim, M.Y & Aylett, R.S (2007). Feel the difference: A Guide with Attitude! In C. Pelachaud et al. (Eds.) *Intelligent virtual agents*, (pp. 317-330), Springer-Verlag Berlin Heidelberg

Louchart, S., Aylett, R. (2007). From Synthetic Characters to Virtual Actors. Proceedings of International Conference in Artificial Intelligence for Interactive Digital Entertainment (pp. 88-91)

Marsella, S., Johnson, W. L., & La Bore, C. M. (2003). Interactive Pedagogical Drama for Health Interventions, *11th International Conference on Artificial Intelligence in Education* Australia

McQuiggan, S., Robison, J., Phillips, R. & Lester, J. (2008), Modeling parallel and reactive empathy in virtual agents: An inductive approach *International Joint Conference on Autonomous Agents and Multi-Agent Systems, Portugal*.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, 19(2), 177–213.

Morishima, Y., Nakajima, H., Brave, S., Yamada, R., Maldonado, H., Nass, C. & Kawaji, S. (2004). The Role of Affect and Sociality in the Agent-Based Collaborative Learning System, *Lecture Notes in Computer Science*, (pp. 265-275).

Mulken, S. V., Andre, E., & Muller, J. (1998). The persona effect: How substantial is it? Paper presented at the HCI-98, Berlin

Mylonas, Ph., Tzouveli, P., & Kollias, S.(2007). E-learning and intelligent content adaptation: an integrated approach International *Journal of Continuing Engineering Education and Life-Long Learning*, 17(4/5) pp. 273-293, Inderscience.

Neji, M. & Ammar, M., (2007). Agent-based Collaborative Affective e-Learning Framework *Electronic Journal of e-Learning*, 5, 123-134

Ortony, A., Clore, g. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*, Cambridge University Press, 1988.

P1484.1 Architecture and Reference Model, IEEE Learning Techonology Standards Committee Retrieved July 2003 from http://ltsc.ieee.org/index.html.

Paliouras G., Papatheodorou C., & Spyropoulos C.D. (2002). Discovering Learner Communities on the Internet Using Unsupervised Machine Learning Techniques Interacting with Computers, Journal of Interacting with Computers 14(6), (pp.761-791), Elsevier.

Pelachaud, C., Carofiglio, V., Poggi, I., De Carolis, B. & De Rosis, F. (2002). Embodied Contextual Agent in Information Delivering Application. *Proceedings of the First international Joint Conference on Autonomous Agents and Multiagent Systems: Part 2*, (pp. 758-765), ACM Press.

Pentland A.S. (2005). Socially Aware Computation and Communication, *Computer*, 38(3), 33-40.

Picard R.W. (2000). Towards computers that recognize and respond to user emotion, *IBM System Journal*, 39 (3), pp.705–719.

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

Prendinger, H., Dohi, H., Wang, H., Mayer, S., & Ishizuka, M.(2004). Empathic embodied interfaces: Addressing users' affective state, Lecture Notes in Computer Science, 3068 (53-64), Springer.

Raouzaiou A, Tsapatsoulis N, Karpouzis K, Kollias S (2002). Parameterized facial expression synthesis based on MPEG-4. EURASIP Journal on Applied Signal Processing (10), 1021–1038

Ryokai, K., Vaucelle, C., & Cassell, J. (2003). Virtual peers as partners in storytelling and literacy learning. *Journal of Computer Assisted Learning*, 19(2), pp.195–208.Blackwell Synergy.