Computing (2016) 98:215–233
DOI 10.1007/s00607-014-0430-9

A hybrid intelligence-aided approach to affect-sensitive e-learning

Jingying Chen · Nan Luo · Yuanyuan Liu · Leyuan Liu · Kun Zhang · Joanna Kolodziej

Received: 1 May 2014 / Accepted: 9 September 2014 / Published online: 1 October 2014

Abstract E-Learning has revolutionized the delivery of learning through the support of rapid advances in Internet technology. Compared with face-to-face traditional classroom education, e-learning lacks interpersonal and emotional interaction between students and teachers. In other words, although a vital factor in learning that influences a human's ability to solve problems, affect has been largely ignored in existing e-learning systems. In this study, we propose a hybrid intelligence-aided approach to affect-sensitive e-learning. A system has been developed that incorporates affect recognition and intervention to improve the learner's learning experience and help the learner become better engaged in the learning process. The system recognizes the learner's affective states using multimodal information via hybrid intelligent approaches, e.g., head pose, eye gaze tracking, facial expression recognition, physiological signal processing and learning progress tracking. The multimodal information gathered is fused based on the proposed affect learning model. The system provides online interventions and adapts the online learning material to the learner's current learning state based on pedagogical strategies. Experimental results show that interest

J. Chen (✉) · N. Luo · Y. Liu · L. Liu · K. Zhang
National Engineering Center for E-Learning, Central China Normal University, Wuhan, China
e-mail: chenjy@mail.ccnu.edu.cn

L. Liu
e-mail: lyliu@mail.ccnu.edu.cn

K. Zhang
e-mail: zhk@mail.ccnu.edu.cn

J. Chen · Y. Liu · L. Liu · K. Zhang
Collaborative and Innovative Center for Educational Technology (CICET), Wuhan, China

J. Kolodziej
Institute of Computer Science, Cracow University of Technology, Kraków, Poland
e-mail: jokoldziej@pk.edu.pl

and confusion are the most frequently occurring states when a learner interacts with a second language learning system and those states are highly related to learning levels (easy versus difficult) and outcomes. Interventions are effective when a learner is disengaged or bored and have been shown to help learners become more engaged in learning.

Keywords E-Learning · Intelligent system · Affect recognition · Affect learning model · Education cloud

Mathematics Subject Classification 68T10

1 Introduction

Most researchers agree that affect is the combination of valence (pleasantness or hedonic value) and arousal (bodily activation), both valence and arousal can be defined as subjective experiences [1]. Valence is a subjective feeling of pleasantness or unpleasantness; arousal is a subjective state of feeling activated or deactivated. Affect plays an important role in human learning. It influences the ability to process information and accurately interpret events. The human brain works as a system with both affective and cognitive functions that are inextricably integrated. Previous research indicates that students can learn more successfully when they feel interested and happy with the learning process [2]. In contrast, they learn less well when frustrated or anxious; affect interferes with the ability to take in information efficiently and deal with it well [2,3]. Affect-sensitive e-learning means an e-learning system can recognize and react to various affective states of a learner and thereby improve the learner's learning experience.

In recent years, e-learning has become an important means of learning [4]. e-learning brings learning to students, not students to learning, thus making it convenient and engaging for students. Also, e-learning can accommodate different learning styles and facilitates learning through a variety of activities. However, e-learning lacks the emotional interaction between students and teachers that face-to-face traditional classroom education accomplishes. For example, a distressed student might be distracted by a sad memory and lose focus on a subject in the classroom. In such a situation, a verbal intervention from the teacher will help the student refocus and redirect their attention to events in the classroom. Most existing e-learning systems pay attention to the capabilities of cognition, while affect is largely ignored [5]. It is desirable for an e-learning system to provide a realistic, engaging and effective online learning experience. Affect recognition and intervention are essential for keeping a learner engaged in the learning process. Although there is a growing recognition of affect as an important factor, there is a lack of formal theory and supporting technology to enable affect in e-learning. Few e-learning systems incorporate affect recognition and intervention. A study exploring the emotional evolution during e-learning is given in [6], where four kinds of physiological data, e.g., heart rate, skin conductance, blood volume pressure and EEG brainwaves [7,8], were used to classify four learner emotions. Their study was conducted with only one subject who was monitored during

20 study sessions (each session lasted 40 min). Nosu and Kurokawa [9] proposed a multi-modal emotion-diagnosis system to support e-learning based on learner facial expressions and biometrical signals. They detected six facial feature points and three biometrical signals (e.g., pulse rate, breathing rate and finger temperature) to infer eight affective states. The recognition rate for 10 e-learning subjects was 74 %, while the rate was 68 % when only facial expression information was used. Their results confirm the effectiveness of multi-modal emotion diagnosis. Since research on affect recognition in e-learning is in an early stage, most researchers use invasive sensors. There has been much progress in the field of affect recognition in e-learning. However, there are no examples of systems that can fully sense natural human communication of emotion and response in a way that rivals that of another person [10] and few systems integrate affective information with an e-learning application to generate pedagogical strategies to help learners when they need it when in a specific affective state.

Hence, it is necessary to investigate the relationship between affect and e-learning and to design an intelligent e-learning system that interacts naturally with learners. One that can recognize and react to various affective states of a learner and thereby improve the learner's learning experience. In this study, we propose a hybrid intelligence-based approach to foster an e-learning experience that is sensitive to affective states of learners. The approach has been implemented via an e-learning system that mimics human affect recognition and enables intervention. Human beings interpret affective states through multiple modalities, including facial expression, speech and behaviors. Our approach therefore utilizes multiple intelligent techniques to comprehensively monitor the affective states of learners.

The hybrid intelligent system recognizes a learner's affective state using multi-modal information, e.g., head pose, eye-gaze tracking, facial expression recognition, physiological signal processing and learning progress tracking. This multimodal information is fused based on the proposed affect learning model. Synergistic effects have been achieved with the following efforts:

1. Develop affect-sensitive functions (e.g., affect recognition, learner's profile and intervention) and establish the e-learning system upon these functions.
2. Propose an affect learning model to analyze a learner's affective state and infer the learner's learning state based on their affective state and learning progress.
3. Investigate methods to elicit and recognize a learner's affective state while learning, using attention estimation, facial expression recognition and physiological signals (e.g., skin conductance) processing.
4. Provide appropriate interventions to help the learner become better engaged in the learning process.

The proposed method can reliably detect a learner's affective state while learning and infer the current learning state. Experiments and results have verified the effectiveness of the proposed method for recognizing affective states. Based on a learner's learning state and the learning objectives stored in the profile, the e-learning system provides online intervention and adapts the online learning material to better match the learner's current learning state, thereby overcoming one of the limitations of e-learning, lack of instant emotional interaction between the learner and tutor that exists in most e-learning experiences today. The proposed system has been designed to

improve the learner's learning experience by recognizing and then responding to the learner's affective state, no matter where and when the learner accesses the system.

The remainder of this paper presents our efforts to tackle a number of research challenges towards affect-sensitive e-learning. Section 2 of this paper discusses existing work and presents the objectives of this study. Section 3 proposes the affect learning model and the framework of the affect-sensitive e-learning system. Section 4 details the approaches used to recognize the various affective states, including head pose estimation, eye gaze tracking, expression recognition, skin conductance processing, learning progress tracking and multimodal fusion. Section 5 describes the experiments and results as well as demonstrates uses of the learning environment. Section 6 concludes the paper with a summary and proposals for future work.

2 Related work

Affect recognition is essential for realizing an affect-sensitive learning system in which the computer recognizes emotions as effectively as a human being can [11]. The computer associates a person's patterns of behavior with affective state information by observing behavior via sensors such as cameras, microphones, and physiological sensors. We highlight representative works in connection with affect recognition.

Lisetti and Nasoz [12] integrated facial expression and autonomic nervous system signals using a camera and a physiological sensor to recognize a user's emotions (e.g., neutral and happy) and adapted an animated interface agent to mirror user emotions. Maat and Pantic [10] described an intelligent system to support affective multimodal human-computer interaction where the user's actions and emotions are modeled and then used to adapt the interaction and support the user in a chosen activity.

Researchers have explored different methods of investigating the relationship between a learner's affective state and the learning process. Graesser et al. [13] investigated the relationship between emotions and learning by tracking the emotions that college students experienced while learning about computer literacy using an animated pedagogical agent, namely AutoTutor. AutoTutor holds a conversation using natural language, with spoken contributions being provided by the learner. However, the researchers' approach required that the students be instructed to make their own judgments about their present affective state rather than automatically detecting them.

Some researchers presented automatic affective state recognition in learning using multisensory information [14,15]. Kapoor et al. [16] developed the automated Learning Companion, which combines information from cameras, a sensing chair, mouse, wireless skin sensor, and task state to detect frustration [17,18]. It used this information to predict when a user needed help.

Research to detect the affective state of different interest levels when a child is solving a puzzle was conducted in [19]. The authors analyzed the facial features captured by a camera, eight different postures from a sensor chair and the task state. The researchers combined the multisensory data in a probabilistic framework and demonstrated that their method achieved much improved recognition accuracy than did classification based on individual channels.

D' Mello et al. [20] developed approaches to automatically sense the four major emotions (confusion, frustration, boredom and flow) in learning on the basis of dialogue history, facial expressions and body posture. Beverly et al. [21] recognized learner affective states while studying mathematics through the use of multisensory data. However, there is still very little understanding as to which emotions are most important in the learning process and how they influence learning. To date, there is no comprehensive, empirically validated theory of emotion that addresses learning. It is believed that simultaneously engaging in both the practice and the theory helps advance both [11].

Significantly different from existing work, this study proposes an affect learning model that recognizes a learner's affective states and provides pedagogical interventions when needed.

3 Affect-sensitive e-learning system overview

To develop an affect-sensitive e-learning system, an affect learning model is proposed to investigate the relationship between affect and learning. This section presents the framework of an affect-sensitive e-learning system based on the affect learning model and the IEEE Reference Model of the Learning Technology Standards Committee (LTSA) [22]. The basic LTSA reference model is extended to include features of affect recognition and interventions in this paper. The learner's affect states are analyzed based on the proposed affect learning model. Hence, the proposed affect learning model is first introduced, then the basic LTSA reference model is described, finally the framework of an affect-sensitive e-learning system based on the affect learning model and the LTSA reference model is presented.

A learner's emotions are modeled to investigate the relationship between affect and learning. The most widely adopted model is OCC [23], which has been established as the standard cognitive appraisal model for emotions. However, this model does not include many of the affective phenomena observed in natural learning situations, such as interest or boredom [11]. Another model, known as Russell's model of affect [24], suggests emotion is the combination of valence and arousal. Valence represents the positive or negative nature of an emotion expressed by the learner, while arousal indicates a person's activation level. During the learning process, the learner's focus of attention is an important factor. For example, if a learner tends to be happy and activated when talking to a friend or doing something unrelated to learning activities, such an affective state is not correlated with good learning outcomes. In contrast, if a learner is excited when solving a problem, this state is highly correlated with good learning outcomes. Hence, we propose to analyze learner affective states based on valence, arousal and attention and then infer the learning state based on the current affective state and learning progress. The proposed affect learning model is given in Table 1 (Note that for confusion or frustration state, the learning state can be positive or negative which depends on the learning progress).

Learning states are positively or negatively correlated with outcomes and intervention is needed when a negative learning state is encountered. For example, if a learner is happily speaking with a friend (i.e., with positive valence and a high arousal

Table 1 The proposed affect learning model

Valence	Arousal	Attention	Affective states	Learning states
Positive	High	On	Interest	Positive
Positive	High	Off	Disengagement	Negative
Positive	Low	On	Satisfaction	Negative
Positive	Low	Off	Disengagement	Negative
Negative	High	On	Confusion or frustration	Positive/negative
Negative	High	Off	Disengagement	Negative
Negative	Low	On	Boredom	Negative
Negative	Low	Off	Boredom	Negative

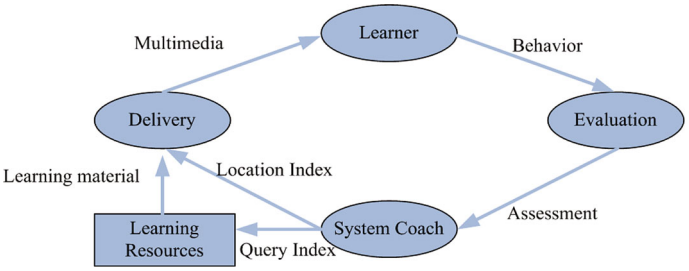

Fig. 1 The basic LTSA model

level but is not focused on learning), the e-learning system may prompt a reminder to prompt the learner to refocus on the learning activity; if a learner shows interest in and is engaged with learning, which is a good state for learning, no intervention is needed. As for the learning-related confusion state, whether intervention is needed or not depends on the learner's progress. For example, intervention is needed if the learner has been stuck for some time without any progress, while no intervention is needed if the learner has made progress.

The reference model of LTSA represents the information flow and linkages between various modules and the interactions between the main processes with the learning value chain. The basic LTSA model is shown in Fig. 1 and 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) transfer information from one entity to another [25]. As illustrated in Fig. 1, a learner's behavior is analyzed using the evaluation process. The system coach then selects the appropriate learning material based on the evaluation outcome. Finally, the delivery process presents the learning material to the learner.

Figure 2 presents the framework of the proposed affect-sensitive e-learning system, which possesses features of affect recognition, the learner's profile and interventions. When the learner interacts with the system, the registration management locates the learner profile according to the learner's ID, which contains a learner's learning objectives and performance (affective states and learning progress). The system coach

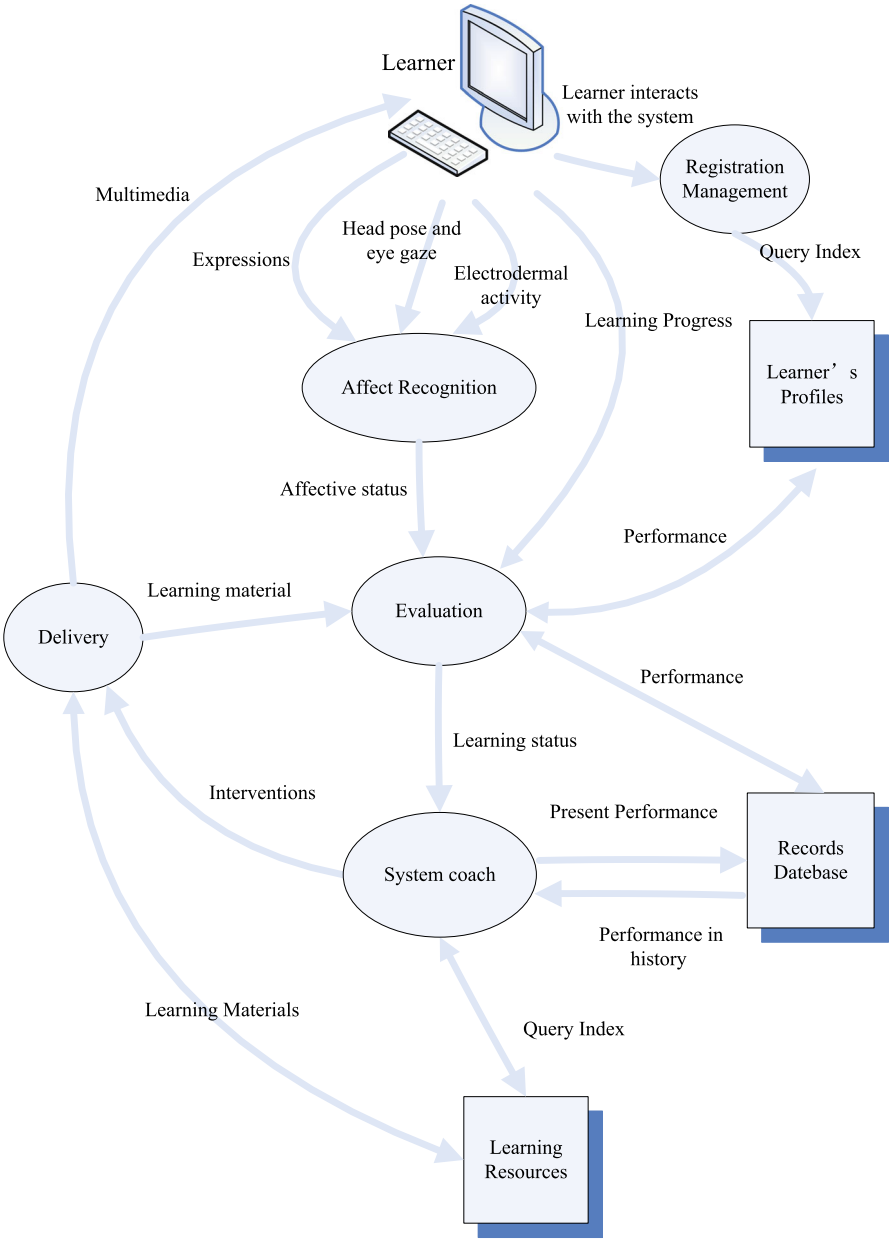

Fig. 2 The framework of the proposed affect-sensitive e-learning system

then locates the learning material from the learning resources which are stored in the web server that best match the learner's profile. In the learning process, the learner's behaviors are captured by cameras and a wireless skin conductance sensor, multimodal information, i.e., head pose, eye gaze, facial expressions and electrodermal activity

are analyzed to infer the affective state based on the proposed affect learning model in the affect recognition process. Then the affective states are input into the evaluation process together with the learning progress. A learner' s each movement and his/her progress (e.g., the time spent answering questions, the number of attempts to answer a question, the number of correct/incorrect answers and the final score) were recorded in a log file, which can be helpful when analyzing learning states. The evaluation process generates the learner's learning state and sends them to the system coach. The evaluation process also sends the learner's performance to the learner's record database and profile for late use. The profile is dynamically updated. The record database stores the learner's present and historic performance. The system coach provides appropriate learning material and interventions for the learner based on analysis of the learner's performance. The delivery process presents the learning material with multimedia and appropriate interventions to the learner. Affect recognition is the key to establish the affect sensitive e-learning system, details on the affect recognition approaches are discussed in Sect. 4.

4 Affective state recognition

To help the system recognize a learner's affective state, a hybrid intelligent framework is proposed to detect the attention, valence, arousal and learning progress of a learner. The framework integrates multimodal data, including facial features detected by cameras and physiological signals captured by a wireless skin conductance sensor. Head pose and eye gaze are estimated to detect a learner's attention, facial expression is classified to predict valence, skin conductance measuring electrodermal activity is processed to correlate with arousal and a log file of a learner's progress is analyzed to track the learning progress.

This section first details the approaches of attention detection, which include face, eyes and mouth detection, facial feature points detection and tracking, head pose and eye gaze estimation. It then presents the expression recognition method, the methods for skin conductance signal processing and learning progress tracking respectively. The section ends with the multimodal fusion method.

4.1 Attention detection

Our approach to attention detection is based on the head pose and eye gaze estimation using multiple facial features. It detects a human face using a boosting algorithm and a set of Haar-like features [26]. The approach then (1) locates the eyes and mouth based on Haar-like features, (2) extracts facial feature points using the extended active shape model (ASM) with the detected eye centers and mouth center as the initial shape position, (3) tracks the detected features using an improved optical flow-based algorithm that incorporates a tracking failure detection mechanism, (4) estimates the head pose using a simple 3D facial feature model, and (5) detects the eye gaze using a linear mapping of vectors between eye corners and pupil centers.

4.1.1 Face detection

Haar-like features can encode the existence of oriented contrast between neighboring regions in an image. Yang [27] demonstrated several types of Haar-like features representing the horizontal, vertical and diagonal intensity information of a facial image at different positions and scales. A cascade of boosted classifiers, i.e., an ensemble of weak classifiers rather than a single strong classifier, working with Haar-like features was trained with a few hundred sample views of face and non-face examples scaled to the same size (24×24 pixels). Thus, simple classifiers at an early stage can efficiently filter out most negative examples and stronger classifiers at the later stage are only necessary to deal with images that look like faces. The trained classifier can be applied to a region of interest of an input image by simply moving the search window across the image to check every location with the classifier.

4.1.2 Eyes and mouth detection

Similar to face detection, eyes can be found using a cascade of boosted tree classifiers with Haar-like features. A statistical model of the eyes is trained. The cascade is trained on 3,000 eye and 8,000 non-eye samples of size 18×12. The training set contains different facial expressions and head rotations. The 18×12 window moves across the eye region and each sub-region is classified as eye or non-eye. Similarly, a cascade of boosted classifiers for the mouth is trained using 2,000 mouth and 4,000 non-mouth samples of size 30×15 in this work. The training set contains different facial expressions and head rotations as well. To search for the mouth, one can move the search window across the lower half of the face image and check every location with the classifier.

4.1.3 Facial feature points detection and tracking

With the detected eye centers and mouth center as the initial points, the extended ASM is used to extract facial feature points. Milborrow and Nicolls [28] improved the original ASM by (1) fitting more landmarks than are actually required, (2) selectively using two- instead of one-dimensional landmark templates, (3) adding noise to the training set, (4) relaxing the shape model where advantageous, (5) trimming covariance matrices by setting most entries to zero and (6) stacking two ASM in series. The extended ASM is efficient and accurate for practical use. However, the extended ASM is sensitive to the initialization, as such accurate positioning of the initial shape is crucial. Our model detects both the eye centers and the mouth center to provide a more accurate initialization. The model is trained with data, including more facial expressions than used in [28] and two subsets of 6 feature points (i.e., two inner eye corners, two outer eye corners, and two nostrils) and 21 feature points (representing the expressive facial features effectively) are used to estimate head pose and classify expressions respectively. Examples of the extracted 21 feature points are shown in Fig. 3.

After feature point detection, an improved optical flow-based algorithm is used to track the feature points. We integrate a tracking failure detection mechanism into the

Fig. 3 Examples of extracted feature points

tracker by calculating the Euclidean distance as:

$$D_{fb} = ||X_t - X_t^{'}||_2 \tag{1}$$

where X_t represents the point in time t, X_{t+1} is the point predicted from X_t by forward optical flow in time $t+1$ and $X_t^{'}$ is the point predicted from X_{t+1} by backward optical flow in time t. If the average distance of the feature points exceeds a certain threshold, it is considered a failure. Once tracking failure has been detected, the feature points must be re-initialized.

4.1.4 Head pose and eye gaze estimation

After the positions of the tracking points have been updated, the head pose can be estimated using POSIT algorithms [29]. Given the 3D facial feature model (i.e., 3D locations of the 6 feature points, the inner and outer eye corners and nostrils, they are robust to non-rigid facial expressions), and their 2D locations in the camera image, the head pose (rotation and translation) with respect to the camera can be computed using the POSIT algorithm. It estimates the pose by first approximating the perspective projection as a scaled orthographic projection and iteratively refines the estimate until the distance between the projected points and those obtained from the estimated pose falls below a certain threshold. Rather than using all of the feature points, a minimal subset of feature points is used to estimate the pose that is valid provided that one subset of good, accurate measurements exists. Once the best subset of the features is determined, the true position of an outlier can be predicted by projecting its model point onto the image using the computed pose. The selection of a good subset can be accomplished within the RANSAC regression paradigm [30].

Based on the estimated head pose, eye gaze can be detected using the inner eye corners noted in the previous section together with the pupil centers. Pupils generally have lower intensities than neighboring pixels and contain a relatively fixed proportion of the information available within the eye regions. In order to compensate for web cameras with low resolution settings and practical use of the environment that is subject to variation in illumination, a robust segmentation approach was developed to extract pupils via an entropy analysis of the partial histogram within the eye regions. Entropy E_j is iteratively calculated according to different parts of the normalized histogram

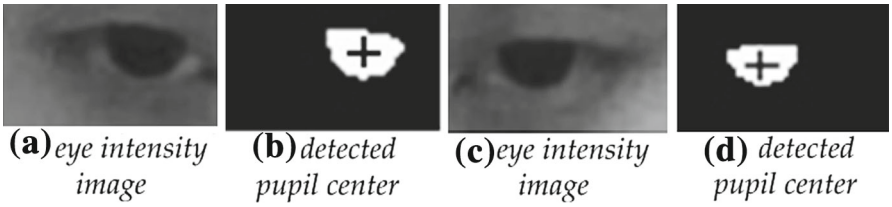

Fig. 4 Detection of pupil centers

until its value is greater than a threshold Eth, which is found from the training data.

$$E_j = -\sum_{i=0}^{j} H(i) \log H(i),\ j = 0, ..., n, n \in (1{,}255) \tag{2}$$

where i and $H(i)$ are histogram index and value respectively. When $E_j > E_{th}$, j is used as a threshold to segment pupils. The centers of the two largest connected bright regions are therefore computed as the centers of the pupils (see Fig. 4). Eye gaze is then estimated by establishing the relationship between gaze and the vector from the inner eye corner to the pupil center.

A calibration procedure is required that requires the user to look at several points on the display so the system can record the positions of each pupil center and inner eye corner as they correspond to a specific gaze. A 2D linear mapping of the vector between the eye corner and the pupil center to the gaze direction is then constructed. Gaze directions can be tracked by interpolation. The proposed eye gaze tracking method uses entropy-based segmentation to ensure accurate tracking, even if the input video has low resolution or captures images under poor conditions.

4.2 Expression recognition

A novel method using both geometric and appearance features of the difference between the normalized neutral image and the fully expressive facial expression image is proposed in this work. The difference tends to emphasize the facial aspects that have changed from the neutral to expressive face and thus eliminates identity of the facial image. The geometric shape features are facial feature point displacements between the normalized neutral and expressive facial expression images, while the appearance features are local texture (i.e., in the neighborhood of facial points) differences between the normalized neutral and expressive facial expression images. Based on the fusion of facial feature point displacements and local texture differences, an SVM-based method is used to recognize facial expressions.

First, the face is normalized based on the estimated head pose. The 21 feature points as described in Sect. 4.1.3 are then located for expression recognition. The displacements between x and y coordinates of 21 feature points on the neutral and expressive faces are then calculated as a 21×2 dimensional geometric feature vector. Texture differences between the normalized neutral and expressive face images are

calculated as texture features. The gradient value is a good measure for describing how the gray level changes within a neighborhood and is less sensitive to light changes. Hence, it can be used to derive the local texture difference measurement. Let $p = \{x, y\}$, $f_n(p)$, $f_e(p)$ and $f'_n(p)$, $f'_e(p)$, be a feature point, the neutral and expressive face images and the neutral and expressive face image gradient vector respectively. Since the normalized cross correlation is robust to noisy conditions, a measure of gradient difference in a feature point's neighborhood between the normalized neutral and expressive face images is defined as:

$$\mathrm{TD} = \frac{\sum_{p \in M} \left(f'_n(p) - \overline{f'_n(p)} \right) \left(f'_e(p) - \overline{f'_e(p)} \right)}{\sqrt{\sum_{p \in M} \left(f'_n(p) - \overline{f'_n(p)} \right)^2 \sum_{p \in M} \left(f'_e(p) - \overline{f'_e(p)} \right)^2}} \tag{3}$$

where $\overline{f'_n(p)}$ and $\overline{f'_e(p)}$ are the averages of $f'_n(p)$ and $f'_e(p)$ and M is the 10 $\times$ 10 neighborhood centered at point p. If there are no large differences within the neighborhood of a feature point p in two images, the local texture features would be similar (i.e., the value of TD is high). If there are large differences in the corresponding regions of two images, there is usually large differences between the local textures of the two images (i.e., the value of TD is low). Texture features are calculated in the neighborhoods centered at the detected 21 feature points.

After the feature point displacements and local texture differences between the normalized neutral and expressive face images have been calculated, the combined feature vector containing a 42 dimensional geometric feature vector and a 21 dimensional texture feature vector is generated for classification. Support vector machine (SVM) is a popular machine learning method for pattern recognition. SVM finds a hyperplane or surface that maximizes the margin between positive and negative observations for a specified class. In this study, we used LIBSVM [31] for training and testing purposes.

4.3 Skin conductance signal processing

Electrodermal activity refers to electrical changes measured at the surface of the skin that arise when the skin receives innervating signals from the brain. For most people, when there is an experience of emotional arousal or increased cognitive workload, the brain sends signals to the skin to increase the level of sweating. One may not feel any sweat on the surface of the skin, but electrical conductance increases in a measurably significant way as the pores begin to fill below the surface. Skin conductance is one form of electrodermal activity. There are two types of skin conductance data: tonic and phasic. Tonic skin conductance is generally considered to be the level of skin conductance in the absence of any particular discrete environmental event or external stimuli. This slow-changing level is generally referred to as skin conductance level (SCL). Phasic skin conductance is typically associated with short-term events and occurs in the presence of discrete environmental stimuli (cognitive processes that precede an event, such as anticipation and decision making). Phasic changes usually

Fig. 5 An example of a skin conductance signal

show up as abrupt increases, or "peaks", in skin conductance. These peaks are generally referred to as skin conductance responses (SCRs). The skin conductance data was collected using a wireless skin conductance bracelet sensor in this study. The raw signal was filtered to remove noise, such as occasional spikes and sudden drops. The baseline of the learner's SCL is measured before learning, while the learner is relaxed. The amplitude of SCRs during learning is analyzed to predict a learner's arousal.

An example of a skin conductance signal is given in Fig. 5. Levenson [32] suggests 0.5–4.0 s as the approximate duration of emotions. We measured the mean of filtered skin conductance data for 4.0 s at 2 Hz (a higher rate is possible, but 2 Hz was sufficient for this study), if the mean is 15 % greater than the baseline, the arousal is assumed to be high, otherwise it is low.

4.4 Learning progress tracking

During the learning process, each movement of a learner and their progress (e.g., the time spent answering questions, the number of attempts to answer a question, the number of correct/incorrect answers, and the final score) were recorded in a log file, which can be helpful when analyzing learning states. For example, if the learner makes several attempts to answer a question, it may be that the learner was confused or under stress and appropriate support was needed; if the learner answers questions correctly and quickly, the learner is assumed to be achieving the learning goals; if the learner cannot answer the question within the specified time, the learner may need intervention (e.g., a prompt to remind the learner to refocus on learning if he has failed to answer the question due to disengagement).

4.5 Multimodal fusion

The fusion of different modalities is generally performed at two levels: feature level and decision level [33,34]. In the feature level fusion, the features extracted from multiple data sources are first combined and then analyzed to make a decision. In decision level fusion, the features of each data source are first analyzed to make an individual decision. The decisions are then combined to make a final decision. Decision level fusion is considered to be the most robust, is resistant to individual sensor, and

is computationally less expensive than feature fusion [35,36]. Another advantage of decision fusion strategy is that it allows use of the most suitable methods for analyzing a single modality. In this study, decisions about head pose, eye gaze, facial expression, skin conductance and the learning log file (i.e., attention, valence, arousal and learning progress) are fused based on the affect learning model. Temporal alignment among different modalities is completed prior to fusion.

We provide a solution to the lack of emotional interaction between learner and tutor in most existing e-learning systems via affective state recognition using the proposed approaches. Therefore, the system can provide online interventions and adapt the online learning material to the learner's state based on pedagogical strategy to better keep the learner engaged in the learning material and process.

5 Experiments and results

Experiments were carried out in a learning lab and included 30 postgraduates (17 females and 13 males). Each student took a pretest survey to evaluate their attitudes toward their second language (English) and their English levels in the same lab. In accordance with their levels, the students were to complete two sessions to learn English for postgraduate english test (e.g., one session was easy and the other was difficult) using the adopted e-learning system over two consecutive days. Each session lasted 50 min and consisted of three segments (i.e., music, vocabulary and reading). A 10-min musical interlude was included to relax the student and obtain baseline measurements of the skin conductance data. The reference value for valence (positive/negative), arousal (high/low) and attention (on/off) from each student was observed by a trained person together with the student's self-report. The data from different modalities were collected by cameras, wireless skin conductance bracelet sensors and log files stored in the system (see Fig. 6).

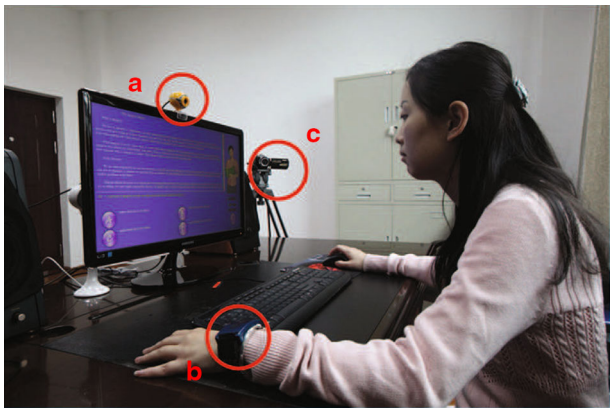

Fig. 6 Learning environment: *a* web camera, *b* skin conductance sensor and *c* video camera

Table 2 Distribution of the affective states during easy and difficult learning sessions

	$V(+)$ and $A(+)$ Interested	$V(+)$ and $A(-)$ Satisfaction	$V(-)$ and $A(+)$ Confusion	$V(-)$ and $A(-)$ Boredom
Easy (%)	50.1	20.2	12.5	17.2
Difficult (%)	23.6	15.3	39.8	21.3

$V(+)$ valence (positive), $V(-)$ valence (negative), $A(+)$ arousal (high), $A(-)$ arousal (low)

5.1 Affective state recognition results

All data was temporally aligned and sampled at the rate of 2 Hz. The head pose and eye gaze determined from a web camera was tracked and used to detect whether the student's gaze was directed at the computer screen or not. The 3D facial feature model of the student was built using a stereo camera before the first learning session. The facial expressions recorded from the same web camera were recognized to predict valence using a classifier, which had been trained from 2,500 face images captured when learners were interacting with the e-learning system. The skin conductance data was first filtered and the mean as the baseline was calculated during the musical interludes. The mean within each 4.0 s interval during the learning session was measured and compared to the baseline to predict the arousal state. Videos captured from the web and video cameras were used for human annotation. The classification rates of attention, valence and arousal were 97.5, 88.3 and 93.5 %, respectively. The classification rates are obtained with 30 subjects learning for two sessions excluding the 10-min musical interlude. From this experiment, we can see that the distributions of the affective states of all students related to learning over two sessions are different (see Table 2).

From this table, we can see that the interest and confusion states occurred most frequently during learning and are highly related to learning outcomes. The confusion state occurs most frequently when the learner found the learning material difficult and achieved a lower score. The interest state occurs most frequently when the learner felt confident and achieved a high score. This latter result is consistent with research findings that a slightly positive mood not only makes a person feel a little better, but works to induce a different kind of thinking that is characterized by a tendency toward greater creativity and flexibility in problem solving, as well as more efficient and thorough decision making [37]. We also found that 90 % of the mean arousal level during the reading segment was higher than during the vocabulary segment. This is consistent with student reports that they concentrate more while reading than when working on vocabulary.

5.2 Intervention studies

According to the detected affective states and recorded learning progress, interventions are generated (see Table 3). The intervention strategy is designed based on the expertise of experienced teachers and experts. An example of the virtual tutor integrated English e-learning interface is given in Fig. 7. The reading material is presented at the top left

Table 3 Interventions for the affective and learning states

Affective state	Learning state	Intervention
Interest	Positive	No
Satisfaction	Positive	No
Confusion	Positive (with learning progress)	No
Confusion	Negative (without learning progress)	The virtual tutor says, "This is difficult; would you like to move to an easier part?" The choice is made by the student
Disengagement	Negative	The virtual tutor says "Are you still there? Please continue your study"
Boredom	Negative (with learning progress)	The virtual tutor says, "Are you getting bored? Would you like to move to a more challenging part?" The choice is made by the student
Boredom	Negative (without learning progress)	The virtual tutor says, "Are you getting bored? Let's move to an easier part." The learning system moves the student to an easier part without further choice by the student

Fig. 7 An example of the virtual tutor integrated English learning interface

of the interface, answer choices to the question are positioned at the bottom left of the interface. A virtual tutor stays at the top right corner of the interface and provides different interventions according to the learner's affective and learning states. For example, if affective states of interest and satisfaction are detected, positive learning states are inferred, hence no intervention is given; if an affective state of confusion is detected, the learning state is decided based on the learning progress (e.g., the student is confused and he/she still chooses an answer to the question and continues the

learning, positive learning state is inferred, on the other hand, the student is confused and does not continue the learning within the limited time, negative learning stated is inferred), no intervention is given with positive learning state detected. Appropriate intervention is provided (e.g., the virtual tutor says, "This is difficult; would you like to move to an easier part?" The choice is made by the student) if an affective state of confusion is detected with negative learning state. Details on the interventions for different affective and learning states are given in Table 3.

After each learning segment, the virtual tutor makes a remark to encourage the student to continue studying, according to the student's learning progress, e.g., "Congratulations, you are doing very well!" or "Those questions are pretty hard, some students need more time to work them out."

The intervention experiment was conducted with the same college students. Each student completed the third 50-min session learning English with interventions. From the experiment, we found that 70 % of the disengaged students were able to refocus on their learning within 2 s after an intervention, 45 % of the confused students chose an easier task and 75 % of the bored students moved to a more challenging part because they found the learning material easy. From these results, it is evident that interventions help students become more engaged in the e-learning system.

As described previously, every student's learning records were stored on the system. The virtual tutor recommends appropriate learning material for a particular student based on learning performance (affective states and learning progress). In this way, the system offers a personalized affect-sensitive learning environment in which a student can learn at his/her own pace.

6 Conclusions and future work

This study established a hybrid intelligence-aided approach to affect-sensitive e-learning. An affect-sensitive e-learning system is described in this paper. We aim to pave the way to an intelligent e-learning system that interacts naturally with learners, can recognize and react to the affective states of a learner and thereby improve a learner's learning experience.

The proposed approach analyzes the learner's affective states via multimodal information, e.g., attention, valence and arousal, via head pose, eye gaze tracking, facial expression recognition and skin conductance signal processing. An affect learning model has been developed to recognize a learner's affective states and infer the learner's learning state based on their current affective state and learning progress. Interventions are provided to improve the learner's learning experience. The learner's performance (affective states and learning progress) are stored in the records database and their profile is updated based on performance relative to material learning outcomes. According to the learner's profile, the system chooses appropriate learning material for the learner.

Through hybrid intelligent methods, an affect-sensitive and personalized e-learning system has been developed to adapt to individual learner needs. This paper presents preliminary results on the system. Affective states can be automatically detected. Experiments and results indicate: (1) interest and confusion are the most frequently

occurring states when learning a second language and are each highly related to the learning level (easy versus difficult) and outcomes, and (2) interventions are effective when a learner is in a disengaged or bored state and can help the learner become better engaged in learning.

In future work, we plan to build a learner model and refine the interventions to deal with negative learning states and thereby improve learning outcomes. Additional evaluation of the effectiveness of interventions will be done. A long-term pedagogical strategy will be generated based on the analysis of the affective states of learners and learning progress to make the e-learning system more effective.

Acknowledgments This work was supported by National Key Technology Research and Development Program (No. 2013BAH72B01) and research funds from Ministry of Education and China Mobile (No. MCM20130601), research funds from the Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJAZH005), research funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU13B001), Wuhan Chenguang Project (No. 2013070104010019), Scientific Research Foundation for the Returned Overseas Chinese Scholars (No. (2013)693),young foundation of Wenhua college (No. J0200540102) and National Natural Science Foundation of China (No. 61272206).

References

1. Russell JA (1989) Measures of emotion. Emotion: theory, research, and experience, vol 4. Academic Press, pp 83–111
2. Bransford JD, Brown AL, Cocking RR (2000) How people learn: brain, mind, experience and school. National Academy Press, Washington, DC
3. Picard RW, Papert S, Bender W et al (2004) Affective learning-a manifesto. BT Technol J 22(4):253–269
4. Luo X, Spaniol M, Wang L, Li Q, Nejdl W, Zhang W (eds) (2010) Advances in web-based learning—ICWL 2010–9th International Conference, Shanghai, December 8–10, 2010. Proceedings Lecture Notes in Computer Science 6483, Springer, ISBN 978-3-642-17406-3
5. Beverly W, Burleson W, Arroyo I et al (2009) Affect-aware tutors: recognising and responding to student affect. Int J Learn Technol 4(3):129–164
6. Shen L, Wang M, Shen R (2009) Affective e-learning: "Emotional" data to improve learning in pervasive learning environment. J Educ Technol Soc 12:176–189
7. Chen D, Li X, Cui D, Wang L, Lu D (2014) Global synchronization measurement of multivariate neural signals with massively parallel nonlinear interdependence analysis. IEEE Trans Neural Syst Rehabil Eng 22(1):33–43
8. Wang L, Chen D, Ranjan R, Ullah Khan S, Kolodziej J, Wang J (2012) Parallel processing of massive EEG data with MapReduce. In: Proceedings of the IEEE 18th international conference on parallel and distributed systems (ICPADS), pp 164–171
9. Nosu K, Kurokawa T (2006) A multi-modal emotion-diagnosis system to support e-learning. In: IEEE proceedings of the first international conference on innovative computing, information and control, pp 274–278
10. Maat L, Pantic M (2007) Gaze-x, adaptive, affective, multimodal interface for single-user office scenarios, artifical Intelligence for Human Computing. Springer, Berlin, pp 251–271
11. Picard RW (2000) Affective computing. MIT Press, Cambridge
12. Lisetti CL, Nasoz F (2002) MAUI: a multimodal affective user interface. In: Proceedings of the tenth ACM international conference on Multimedia. ACM, pp 161–170
13. Graesser A, Chipman P, King B, et al (2007) Emotions and learning with auto tutor. In: Proceedings of the conference on artificial intelligence in education: building technology rich learning contexts that work, pp 569–571
14. Xia F, Yang LT, Wang L, Vinel AV (2012) Internet of things. Int J Commun Syst 25(9):1101–1102
15. Chang Y, Jung J, Wang L (2014) From ubiquitous sensing to cloud computing: technologies and applications. Int J Distrib Sens Netw

16. Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int J Hum Comput Stud 65(8):724–736
17. Kolodziej J, Ullah Khan S, Wang L, Min-Allah N, Ahmad Madani S, Ghani N, Li H (2011) An application of Markov jump process model for activity-based indoor mobility prediction in wireless networks. FIT, pp 51–56
18. Jie W, Cai W, Wang L, Procter R (2007) A secure information service for monitoring large scale grids. Parallel Comput 33(7–8):572–591
19. Kapoor A, Picard RW, Ivanov Y (2004) Probabilistic combination of multiple modalities to detect interest. In: Proceedings of the 17th IEEE International Conference on Pattern Recognition, vol 3. pp 969–972
20. D'Mello S, Picard R, Graesser A (2007) Towards an affect-sensitive autotutor. IEEE Intell Syst 22(4):53–61
21. Beverly W, Burleson W, Arroyo I, Dragon T, Cooper D, Picard R (2009) Affect-aware tutors: recognizing and responding to student affect. J Learn Technol 4:129–164
22. IEEE Learning Technology Standards Committee (2001) Draft Standard for Learning Technology Learning Technology Systems Architecture (LTSA), IEEE Computer Society. http://www.computer.org/portal/web/sab/learning-technology
23. Ortony A (1990) The cognitive structure of emotions. Cambridge University Press, Cambridge
24. Russell JA (1980) A circumplex model of affect. J Personal Soc Psychol 39(6):1161
25. Caridakis G, Tzouveli P, Raouzaiou A, Karpouzis K, Kollias S (2010) Affective e-learning system: analysis of learners state, affective, interactive and cognitive methods for e-learning design: creating an optimal education experience
26. Viola P, Jones M (2001) Robust real-time object detection. Int J Comput Vis 4:34–47
27. Yang M (2009) Face detection. In: Stan ZL (ed) Encyclopedia of biometrics. Springer
28. Milborrow S, Nicolls F (2008) Locating facial features with an extended active shape model. In: Proceedings of the 10th european conference on computer vision: Part IV
29. Dementhon DF, Davis LS (1995) Model based object pose in 25 Lines of Code. Int J Comput Vis 15:123–141
30. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. J Commun ACM 24:381–395
31. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27
32. Levenson RW (1988) Emotion and the autonomic nervous system: a prospectus for research on autonomic specificity, Soc Psychophysiol Emot Theo Clin Appl. pp 17–42
33. Wei J, Liu D, Wang L (2014) A general metric and parallel framework for adaptive image fusion in clusters. Concurr Comput Pract Exp 26(7):1375–1387
34. Atrey PK, Hossain MA, El Saddik A et al (2010) Multimodal fusion for multimedia analysis: a survey. Multimed Syst 16(6):345–379
35. Sharma R, Pavlovic VI, Huang TS (1998) Toward multimodal human-computer interface. Proc IEEE 86(5):853–869
36. Pantic M, Rothkrantz LJM (2003) Toward an affect-sensitive multimodal human-computer interaction. Proc IEEE 91(9):1370–1390
37. Isen A (2000) Positive affect and decision making. Handbook of emotios, Guilford