We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

6

Multimodal Intelligent Tutoring Systems

Xia Mao and Zheng Li
School of electronic and information engineering, Beihang University, Beijng China

1. Introduction

Intelligent Tutoring Systems (ITS), also known as Intelligent E-learning Systems, is the task of providing individualized instruction, by being able to adapt to the knowledge, learning abilities and needs of each individual student. The offer of the ITS is increasing at an unrestrainable pace (Butz & Hua, 2006; Chen, 2008; Reategui & Boff, 2008; Roger, 2006). The "intelligence" in these systems is seen through the way they adapt themselves to the characteristics of the students, such as speed of learning, specific areas in which the student excels as well as falls behind, and rate of learning as more knowledge is learned. However, they are still not as effective as one-on-one human tutoring. We believe that an important factor in the success of human one-on-one tutoring is the tutor's ability to identify and respond to the student's attention information and affective state. In ITS, these two characteristics may give the student the sensation that there is someone behind the program who follows his learning development and cares about him as a human tutor would. In this paper, we propose the Multimodal Intelligent Tutoring Systems (MITS), which detects the attention information and affective state and uses the information to drive the agent tutor to individualize interaction with the learner.

2. Architecture for MITS

ITS is a computer-based educational system that provides individualized instruction like a human tutor. A traditional ITS decides how and what to teach based on the student pedagogical state. However, it has been demonstrated that an experienced human tutor can manage the attention information and affective state (besides the pedagogical state) of the student to motivate him and to improve the learning process. Therefore, the interface between the student and tutor in traditional ITS needs to be augmented to include attention information interface and affective information interface. ITS needs the ability of reasoning about the attention information and affective state to provide students with an adequate response from a pedagogical, attentive and affective point of view; in this sense, the module of attention information processing and affective information processing are required. The attention information processing module analyzes the gaze behavior of student in real-time and is capable of adapting the presentation flow according to the student's interest or non-interest. The affective information processing module analyzes the facial expression, speech and text input by the student to sense the underpinned affective qualities. Once the attention information and affective state have been obtained, the agent tutor has to respond accordingly. We must enable a mapping from the attention information and affective state

to actions of the agent tutor. We refined our tutoring strategy module by means of questionnaires presented to teachers. In the questionnaires we presented several scenarios of tutoring and asked the teachers to give the appropriate pedagogical and affective action for each scenario. The affective action includes the facial expression, emotional speech synthesis and text that produced from the Artificial Intelligence Markup Language (AIML) Retrieval Mechanism. The Architecture of MITS can be seen in figure 1.

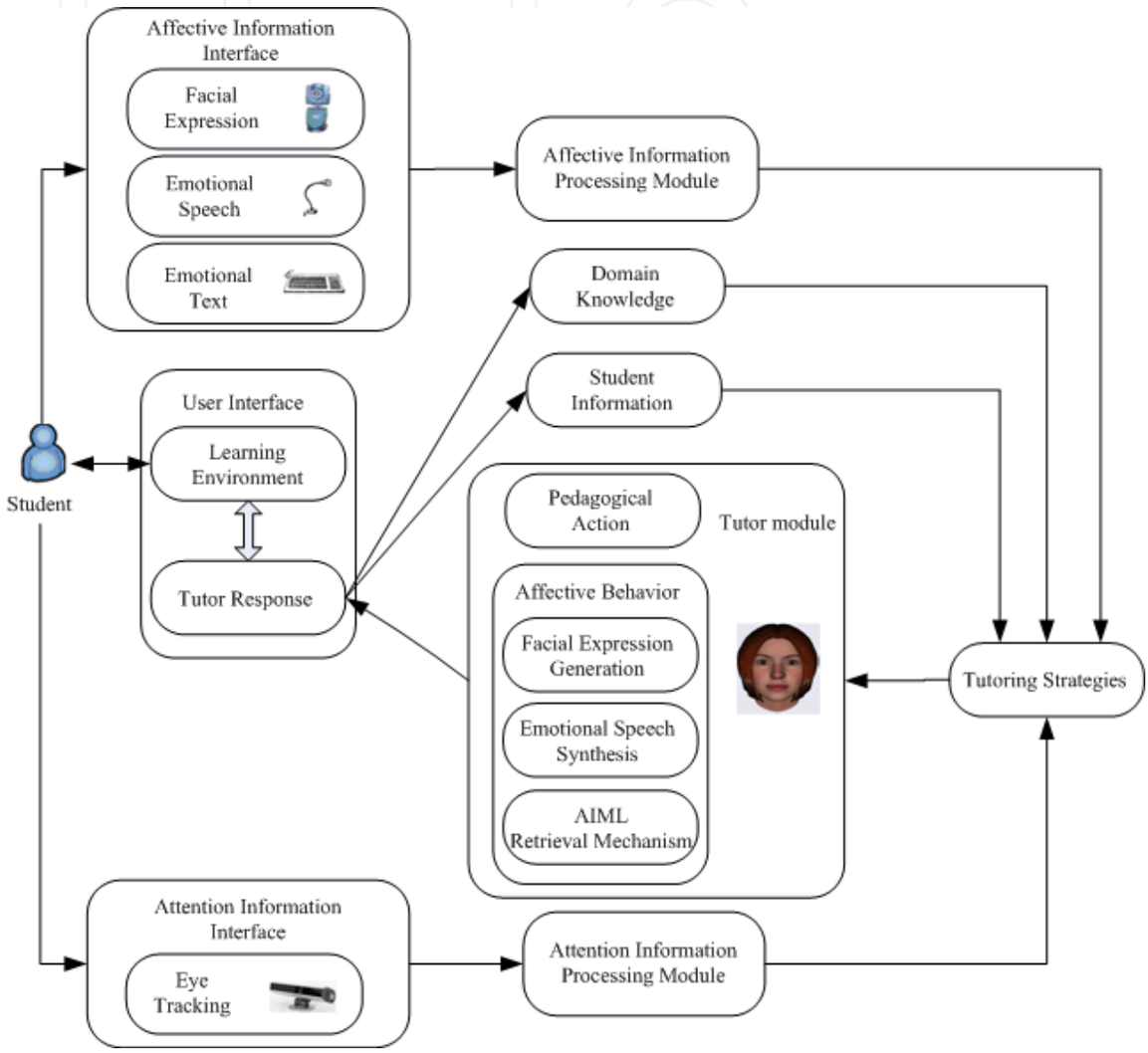


Fig. 1. Architecture of MITS.

3. Attention information

Being "a window to the mind", the eye and its movement are tightly coupled with human cognitive processes. In this paper, we use the eye tracker iView RED from SMI (IView RED. http://www.smivision.com/) to follow student's gaze. Eye movement provides an indication of student's interest and focus of attention. Screen areas that may trigger a system response when being looked at (or not looked at) are called "interest areas". Figure 2

illustrates one example of the interest areas. For each interest area, the interest score is calculated. When the score for an area exceeds a threshold, the agent will react if a reaction is defined.

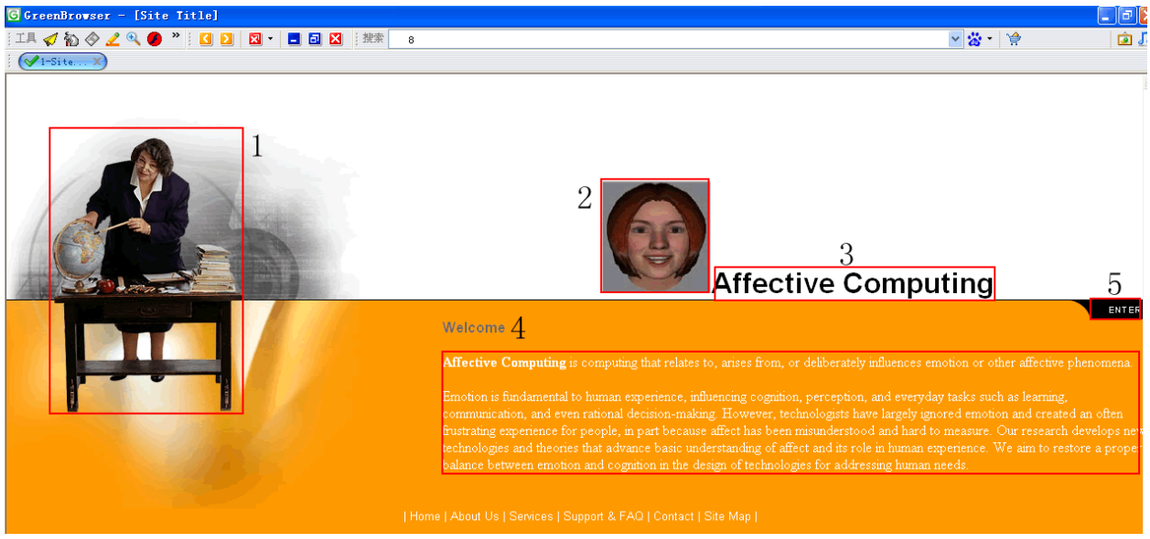


Fig. 2. Example of "interest areas".

The key functionality of the attention information processing module in our MITS is characterized by three main components:

- Monitor the grounding. In human face-to-face communication, grounding relates to the process of ensuring that what has been said is understood by the conversational partners, i.e. there is "common ground". During the learner-tutor interaction, grounding is considered successful if the following condition is met: the student's gaze shows a transition from the screen area of the speaking tutor to the screen area of the referent mentioned by the tutor. When positive evidence in grounding is observed, the course will continue. And a window of contextual content, i.e. the related contend, maybe popup according to the referent.
- Guarantee the attention. The agent will perform an interruption if the student attends to interest areas that are not considered as part of the current content the agent is talking about. An "alert" action will be performed if the student does not gaze at the display, for example, gaze out of the window.
- Note the history. This function records which area and how much of the area has been accessed by the student. If an important area that the student does not pay enough attention to, this area might be proposed again. While the area previously has been accessed for enough time, it is not very likely that the student intends to activate it again.

The components aforesaid are all based on the modified version of the algorithm described by Qvarfordt (Qvarfordt & Zhai, 2005), where it is used for an intelligent virtual tourist information environment (iTourist). Two interest metrics were developed: (1) the Interest Score (IScore) and (2) the Focus of Interest Score (FIScore). IScore is used to determine an area's "arousal" level, or the likelihood that the user is interested in it. When the IScore metric passes a certain threshold, the area is said to become "active". FIScore measures how

the user keeps up his or her interest in an active area. If the FIScore for an active area falls below a certain threshold, it becomes deactivated and a new active area is selected based on the IScore. According to the key functionality of the attention information processing module in our system, a simplified version of the IScore metric is sufficient for our purpose. IScore basic component is eye-gaze intensity p :

$$p = \frac{T_{ISon}}{T_{IS}} \tag{1}$$

Where T_{ISon} refers to the accumulated gaze duration within a time window of size T_{IS} (in our system, 1000 ms) and T_{IS} is the size of the moving time window. In order to account for factors that may relate to user's interest, Qvarfordt characterized the IScore as $p_{is} = p(1+\alpha(1-p))$, where p_{is} is the arousal level of the area and α is the excitability modification defined as below in (Qvarfordt & Zhai, 2005).

$$\alpha = \frac{c_f\alpha_f + c_c\alpha_c + c_s\alpha_s + c_a\alpha_a}{c_f + c_c + c_s + c_a} \tag{2}$$

Where α_f , α_c , α_s , α_a are constants empirically adjusted, they are defined as:

- α_f is the frequency of the user's eye gaze entering and leaving the area
- α_c is the categorical relationship with the previous active area
- α_s is the relative size to a baseline area
- α_a records previous activation of the area

We modified the formula 2, only α_f , α_s , α_a were integrated into MITS. The factor α_f is represented as $\alpha_f = \frac{N_{sw}}{N_f}$, where N_{sw} denotes the number of times eye gaze enters and leaves the area and N_f denotes the maximum possible N_{sw} in the preset time window. α_f is identified as one indication of a user's interest in an area. Since some noise in the eye movement signal, larger areas could have a higher chance of being "hit" than smaller ones, α_s is defined to avoid this. α_s is represented by $\alpha_s = \frac{S_b - S}{S}$, where S_b is the area size of the common areas which are also the smallest, and S represents the size of the current area. As for the α_a , it is employed to indicate whether the area has been paid enough attention. $\alpha_a = -1$ when the area has been paid enough attention and 0 when it has not been paid enough attention.

4. Affective information

Our interest in the emotion integrated in tutoring systems is motivated by the social cognitive theory suggesting that learning takes place through a complex interplay between both cognitive and affective dimensions. Researches in cognitive sciences argue that emotion enables people to communicate efficiently by monitoring and regulating social interaction, by evaluating and modifying emotional experiences. ITS would be significantly enhanced if computers could adapt according to the affective state of the student. In order to

get an idea about the effectiveness of machine-based emotion recognition compared to humans, a review of research has been done by Huang (Huang & Chen, 1998). They investigated the performance of machine based emotion employing both video and audio information. Their work was based on human performance results reported by DeSilva (DeSilva & Miyasato, 1997). Their research indicated that the machine performance was on average better than human performance with 75% accuracy. In addition, comparing detection of confusions indicated similarities between machine and human. These results are encouraging in the context of our research for integrating multimodal affective interaction into tutoring systems. Although the term "affective tutoring systems" can be traced back as far as Picard's book "Affective computing" in 1997, to date, only few projects have explicitly considered emotion for ITS. However, all the projects in existence are single-channel, and mainly concentrated on the facial expression recognition. In this paper, we detect the student's emotion through facial expression, speech and text which are main carriers of human emotion. The following subsection will give a brief description of our methods to capture the emotion through the three channels.

4.1 Facial expression

Facial expression recognition has attracted a significant interest in the scientific community due to its importance for human centered interfaces. Many researchers have integrated the facial expressions into ITS (Reategui & Boff, 2008; Roger, 2006; Sarrafzadeh & Alexander, 2008). However, the performance of facial expression recognition could be influenced by occlusion on the face caused by pose variation, glass wearing, and hair or hand covering etc. The ability to handle occluded facial features is most important for achieving robustness of facial expression recognition. In contrast to normal methods that do not deal with the occlusion regions separately, our approach detects and eliminates the facial occlusions for robust facial expression recognition. Thus, the procedure of facial occlusion removal is added to normal classification procedure. Here, we propose a novel method for partial occlusion removal by iterative operation of facial occlusion detection and reconstruction using RPCA and saliency detection until no occlusion is detected. Then, the reconstructed face after occlusion removal is put to AdaBoost classifier for robust facial expression recognition, as shown in Figure 3.

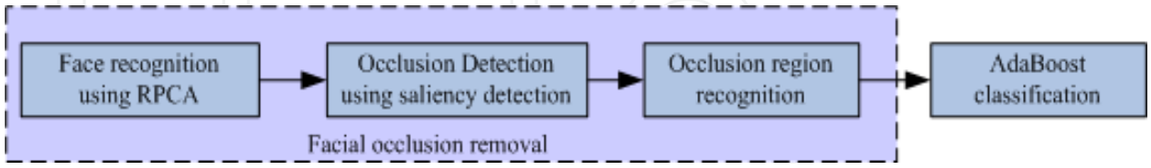


Fig. 3. Work flow of the robust facial expression recognition.

4.1.1 Face recognition using RPCA

Robust principal component analysis (RPCA) is robust to outliers (i.e. artifacts due to occlusion, illumination, image noise etc.) in training data and can be used to construct low-dimensional linear-subspace representations from noisy data. When the face contains a small fraction of the subjects with glasses, or forehead overlaid by hair, or chin surrounded by hands, the pixels corresponding to those coverings are likely to be treated as outliers by RPCA. Hence, the reconstructed image of the original face will possibly not contain the occlusions.

4.1.2 Occlusion detection using saliency detection

To find the occlusion regions on the face, we adopt the method of saliency detection. Firstly, the original face image is transformed into gray level and normalized to $I(x,y)$ using histogram equalization. Then, $I(x,y)$ is reconstructed to $R(x,y)$ using RPCA. We can obtain the residual image $D(x,y)$ between the reconstructed image $R(x,y)$ and $I(x,y)$ by:

$$D(x,y) = \left| R(x,y) - I(x,y) \right| \tag{3}$$

Then, the residual image $D(x,y)$ is put to a saliency detector to find the local places with high complexity, which is hypothesized to be the occlusion on the face. The measure of the local saliency is defined as:

$$H_{D,Rx} = -\sum_i P_{D,R_X}(d_i) \log_2 P_{D,R_X}(d_i) \tag{4}$$

Where $P_{D,R_X}(d_i)$ is the probability of descriptor (or difference image) D taking the value d_i in the local region R_x. We apply the saliency detection in the residual image over a wide range of scale, and set the threshold value of $H_{D,Rx}$. The region with biggest $H_{D,Rx}$ value over the threshold is set to the occlusion region. If all regions have $H_{D,Rx}$ less than the threshold, it is presumed that no occlusion exists. Note that we just choose one occlusion region in one operation of saliency detection even if there are multiple regions with saliency value over the threshold.

4.1.3 Occlusion region reconstruction

Detailed information is most important to facial expression recognition. To avoid the wrong information introduced by face reconstruction in non-occluded region, we adopt the mechanism of occlusion region reconstruction rather than the total face reconstruction. To obtain the new face image $P(x,y)$, pixel values of the detected occlusion region will be replaced by the reconstructed face using RPCA. Thus, the wrong information in the occlusion region may be shielded while the other regions of the face retain the same. To further decrease the impact of occlusion for facial expression reconstruction, we perform occlusion region reconstruction for several iterations until the difference of the reconstructed face between two iterations is below a threshold. The new face image $P_t(x,y)$ in iteration *t* can be obtained by:

$$P_t(x,y) = \begin{cases} I(x,y) & (x,y) \notin R_{occlusion} \\ R_t(x,y) & (x,y) \in R_{occlusion} \end{cases} \tag{5}$$

Where $I(x,y)$ is the normalized image, $R_t(x,y)$ is the reconstructed image using RPCA in iteration *t*, and $R_{occlusion}$ defines the occlusion region. Note that

$$R_t(x,y) = \begin{cases} RPCA(I) & t = 1 \\ RPCA(P_{t-1}) & t > 1 \end{cases} \tag{6}$$

Where *RPCA* designates the RPCA procedure, *t* is the iteration index.

4.1.4 AdaBoost classification

We employ harr-like features for feature extraction and implement multiple one-against-rest two-class AdaBoost classifiers for robust facial expression recognition. In the algorithm, multiple two-class classifiers are constructed from weak features which are selected to discriminate one class from the others. It can solve the problem that weak features to discriminate multiple classes are hard to be selected in traditional multi-class AdaBoost algorithm. The proposed algorithms were trained and tested on our Facial Expression Database. This database consists of 57 university students in age from 19 to 27 years old and includes videos with hand and glass occlusion when displaying kinds of facial expressions. We also randomly add occlusions on the face to generate occluded faces. The experiment results are listed in Table 1.

Emotion	anger	happiness	sadness
Accuracy	83.5	85.3	70.6
Emotion	disgust	surprise	average
Accuracy	75.0	73.3	77.5

Table 1. Facial Recognition Results.

4.2 Speech emotion

In the student-tutor interaction, human tutors respond to both what a student says and to how the student says it. However, most tutorial dialogue systems can not detect the student emotion and attitudes underlying an utterance. In this paper, we introduce the speech emotion recognition into the ITS.

4.2.1 Feature extraction and relative feature calculation

Study on emotion of speech indicates that pitch, energy, duration, formant, Mel prediction cepstrum coefficient (MPCC) and linear prediction cepstrum coefficient (LPCC) are effective absolute features to distinguish certain emotions. In the paper, for each frame, six basic features, including pitch, amplitude energy, box-dimension, zero cross ratio, energy-frequency-value, first formant frequency, as well as their first and second derivatives, are extracted. Besides, 10-order LPCC and 12-order MFCC are also be extracted. Though absolute features of speeches corresponding to same emotion have large differences among different speakers, the differences of feature change induced by emotion stimulation are small relatively. Hence, relative features which reflects feature change is more credible than absolute features for emotion recognition. Relative features used in the paper embody alterations of pitch, energy or other features. They are obtained by computing the change rate relative to natural speech. Features of the kind are robust to different speakers because its calculation is combined with normalization of the features of neutral speeches. For computing relative features, the reference features of neutral version of each text and each speaker should be obtained by calculating the statistics of some frame-based parameters. In this paper, the statistic features used are means of dynamic features, including pitch, amplitude energy, energy-frequency-value, box-dimension, zero cross ratio, and first formant frequency as well as their first and second derivatives. Then, the six statistic features are used to normalize the corresponding dynamic features for each emotion speech, including training samples and test samples. Assuming Mf_i, $i = 1,2,\cdots,18$ are reference features of neutral version, $\vec{f_i}, i = 1,2,\cdots 18$ are the corresponding dynamic feature vectors, the relative feature vectors $R\vec{f_i}$ can be obtained according to following formula:

$$R\vec{f}_i = (\vec{f}_i - Mf_i) / (Mf_i + 0.0000001) \quad (7)$$

where $\vec{f}_i = [f_{i1}, f_{i2}, ...f_{iL}]^T$, $R\vec{f}_i = [Rf_{i1}, Rf_{i2}, \cdots Rf_{iL}]^T$ and L indicates the length of feature vector.

4.2.2 Isolated HMMs

The HMMs are left-right discrete models. The most pervasive methods, Forward-Backward Procedure, Viterbi Algorithm and Baum Welch re-estimation are employed in this paper. Baum Welch re-estimation based on likelihood training criterion is used to train the HMMs, each HMM modeling one emotion; Forward-Backward Procedure exports the likelihood probability; Viterbi Algorithm, focusing on the best path through the model, evaluates the likelihood of the best match between the given speech observations and the given HMMs, then achieves the "optimal" state sequences. The recognizing process based on HMMs is shown as Figure 4. A speech sample is analyzed and then represented by a feature vector, according to which the likelihood between the speech sample and each HMM is computed. Then the emotion state corresponding to maximum likelihood is selected as the output of the classifier through comparison.

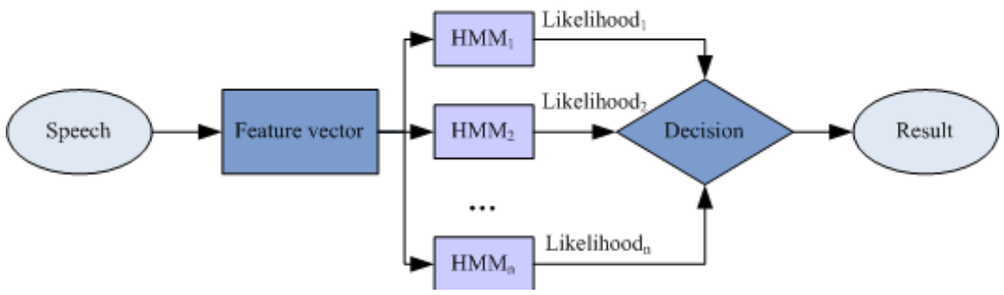


Fig. 4. Emotion Recognition by HMMs.

4.2.3 HMMs fusion system

For the complexity of speech emotion recognition, single classifier systems have limited performance. In recent years, classifier fusion proves to be effective and efficient. By taking advantage of complementary information provided by the constituent classifiers, classifier fusion offers improved performance. Classifier fusion can be done at two different levels, namely, score level and decision level. In score level fusion, raw outputs (scores or confidence levels) of the individual classifiers are combined in a certain way to reach a global decision. The combination can be performed either simply using the sum rule or averaged sum rule, or more sophisticatedly, using another classifier. Decision level fusion, on the other hand, arrives at the final classification decision by combining the decisions of individual classifiers. Voting is a well-known technique for decision-level fusion. It can mask errors from one or more classifiers and make the system more robust. Voting strategies include: majority, weighted voting, plurality, instance runoff voting, threshold voting, and the more general weighted k-out-of-n systems. In this paper, four HMMs classifiers, which have different feature vectors (see Table2), are used. HMMs classifier takes only the emotion which satisfies the model most as the recognition result. But the correct result often should be the emotion which satisfies the model secondly or thirdly. So a new algorithm named weighted ranked voting, which is a reformed version of ranked voting

method provided by C De Borda, is proposed. Ranked voting method permits a voter to choose more than one candidate in proper order. Moreover, the improved algorithm also makes the voted emotions attached by different weights.

classifier	Feature vector
1	pitch, box- dimension, energy with their first and second derivatives; 10-order LPCC
2	energy-frequency-value, box- dimension, formant with their first and second derivatives; 12-order MFCC
3	pitch, zero cross ratio, formant with their first and second derivatives; 12-order MFCC
4	all features extracted in the paper

Table 2. Feature vector.

For a speech sample and a classifier, the voting weight of a certain emotion is determined according to the likelihood between the speech and the HMM model corresponding to the emotion. Firstly, the likelihood values between the speech sample and HMM models are calculated. Secondly, the emotion states are sorted according to likelihood. Then, the voting weights of the first three emotions are allocated according to the order. In the paper, the weight is determined as Table 3. Finally, the weights from four classifiers corresponding to each emotion are summed up and the emotion which has maximum value is selected as result.

	First	Second	Third
Weight	1	0.6	0.3

Table 3. Weight Allocation for Voting.

The steps are listed as follows for each speech sample.

- step1: Initialize weight value as 0 for each emotion.
- step2: Sort emotions according to likelihood for each classifier.
- step3: Vote the first three emotion attached by weight according to Table 3 for each classifier.
- step4: Sum up the weights from four classifiers for each emotion and choose the emotion which has the biggest weight sum as the recognition result.

To evaluate the performance of the proposed classifier in this paper, Database of Emotional Speech was set up to provide speech samples. This corpus contains utterances of five emotions, twenty texts and five actors, two males and three females. Each speaker repeats each text three times in each emotion, meaning that sixty utterances per emotion. For classifier evaluation, 1,140 samples of eight speakers, which have been assessed, are used. The evaluation was done in a "leave-one-speaker-out" manner. One feature vector, formed by six relative features combined with LPCC or MFCC, is used. The experiment results are listed in Table 4.

4.3 Text

Text is an important modality for learner-tutor interaction, many of the ITS have the function enabling the tutor to chat with the student or assist the student in theoretical questions. so studying the relationship between natural language and affective information as well as assessing the underpinned affective qualities of natural language is also

Emotion	Classifiers				
	1	2	3	4	fusion
anger	14.3	19.1	19.1	19.1	14.3
happiness	31.8	45.5	50.0	40.9	100
sadness	8.7	47.8	73.9	47.8	60.9
disgust	92.3	45.5	50.0	40.9	100
surprise	38.9	33.3	33.3	33.3	83.3
average	30.2	37.1	40.5	36.2	57.8

Table 4. Results Using Relative Feature Vector.

important. The Artificial Intelligence Markup Language (AIML) is used to represent the tutor's conversational knowledge, employing a mechanism of stimulus-response. The stimuli (sentences and fragments which may be used to question the tutor) are stored and used to search for pre-defined replies. When the learner poses a question, the tutor starts the AIML Retrieval Mechanism in order to build an appropriate reply using the information, patterns and templates from the AIML database. AIML is an Extensible Markup Language (XML) derivative, which power lies in three basic aspects: AIML syntax enables the semantic content of a question to be extracted easily so that the appropriate answer can be given quickly. The use of labels to combine answers lends greater variety to the answers and increases the number of questions to which an answer can be given. The use of recursivity enables answers to be provided to inputs for which, in theory, there is no direct answer.

Reategui have adopted the AIML in their ITS (Reategui & Boff, 2008), however, it can not sense the affective information conveyed by text automatically. In this paper, we integrate the textual affect sensing algorithm into the AIML Retrieval Mechanism. Figure 5 shows the work flow of the textual affect sensing. The approach for providing emotional estimations from the sentence input by the student is based on a keyword spotting technique and sentence-level processing technique.

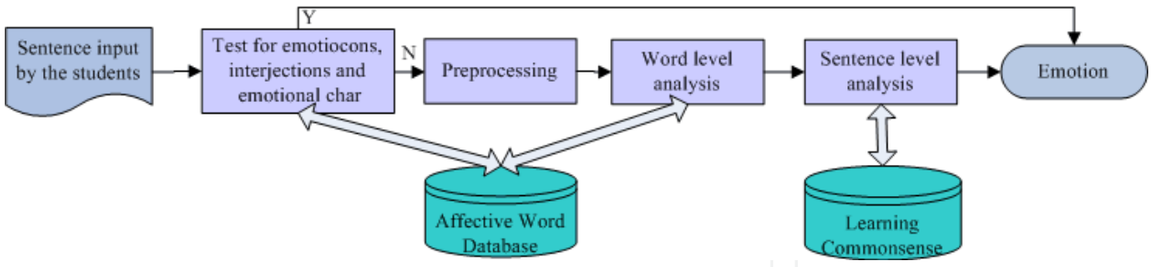


Fig. 5. Work flow of the textual affect sensing.

4.3.1 Affective database and learning commonsense database

In order to support the handling of abbreviated language and the interpretation of affective features of emoticons, abbreviations, interjections and words, an affective database was created using XML. We collected emoticons (such as, ":-)" for happiness and "QQ" for sadness), the most popular emotional acronyms and abbreviations (for instance, "LOL" (laughing out loud) for happiness) and emotional interjections (for example: "damn" for anger and "wow" for happiness). We also have taken emotional adjectives, nouns, verbs, and adverbs words into our database.

Besides the affective database, the learning commonsense database is also constructed. Our idea relies on having broad knowledge about student's common affective attitudes toward learning process. For instance, if the student input "The content is too difficult to understand", it implies that the student is not happy, as for the input "I got a high score in the test" indicates the student is happy. The structure of the learning commonsense database is based on the affect models generated from Open Mind Common Sense (OMCS) (Liu & Lieberman, 2003).

4.3.2 Textual emotion sensing

Firstly, the multiple-sentence input by the student is spited into single sentences. Each sentence is estimating the emotion separately. The sentence is tested for occurrences of emoticons, abbreviations, acronyms, interjections. If there is an emoticon, abbreviation, acronym or interjection related to an emotional state, no further analysis of affect in sentence is performed based on the assumption that the emoticon, abbreviation, acronym or abbreviation dominates the affective meaning of the entire sentence. If there are no emotion-relevant emoticons, abbreviations, acronym or interjection in a sentence, we prepare the sentence for the next processing: we use deep syntactical parser, Connexor Machinese Syntax, returns exhaustive information for analyzed sentences. From the parser output in XML style, we can read off the characteristics of each token and the relations between them in a sentence, such as subject, verb, object, and their attributes. Then, we use the word spotting technique to estimate emotion of word based on the affective database. However, the word spotting method is too simple to deal with sentences without any affective word. We hence perform the following steps on sentence-level processing. In this stage, we search the learning commonsense database to get the emotion effect of the verb. Finally, we detect "negation" in sentences. Since negatively prefixed words such as "unhappy" are already included in the emotion database, they do not have to be considered. On the other hand, negative verb forms such as "was not", "did not" are detected and flip the polarity of the emotion word.

When student inputs sentences, the function of textual affect sensing is called firstly. Then the AIML Retrieval Mechanism (www.alicebot.org/aiml.html) starts in order to generate an appropriate reply using the pattern and template from the AIML database. For instance, if the student input "What is Affective Computing? It sounds really interest!", the pattern with happy is mapped. While the question is "What is Affective Computing? It is really too abstract to understand! Can you help me?", the pattern with sad takes effect. Different answers are retrieved for the two patterns, as shown in the examples below:

<*pattern*>What is Affective Computing HAPPY </*pattern*>
<*template*> Affective Computing is a very interesting topic! It is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena. </*template*>
<*pattern*> What is Affective Computing SAD </*pattern*>
<*template*> Oh, you seem a little unhappy. Be patient and it is easy to understand! Affective Computing is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena. </*template*>

5. Agent tutor

In our MITS, an agent tutor "Alice" can adjust her behavior in response to learner's requests and inferred learner's needs. The agent is "eye-aware" and "affect-aware", and provides consistent empathy using facial expression and synthetic emotional speech. Its emotional response depends on the learner's action. For instance, an agent shows a happy emotion if the learner concentrates on the current study topic. In contrast, if the learner seems to lose

concentration, the agent will show mild anger or alert the learner. The agent also shows empathy when the learner is sad. In general, the agent tutor interacts between the educational content and the learner. Other tasks of an agent tutor include explaining the study material and providing hints when necessary, moving around the screen to get or direct user attention, and to highlight information. The detailed tutoring strategies will be given latter. In this section, we focus on the facial expression generation and emotional speech synthesis of the agent. The famous agent "Alice" is employed as the tutor. Other agent systems can be used with appropriate diver programs.

5.1 Facial expression generation

Facial expression plays an important role in human's daily life, as indicated by Mehrabian, in face-to-face human communication 55% of the communicative message is transferred by facial expressions (Mehrabian, 1968). However, the limit in the existence researches is that facial expression generation is mostly monotone, or in the "Invariable View". They usually correlate one model of facial expression to one emotion, and generate facial animation based on that. Whereas, human tend to act more complicated to express one emotion. For example, human display kinds of facial expressions to express happiness, such as smile with mouth open or closed, symmetrically or asymmetrically, even with head wobbled. In this paper, we aim at generating humanoid and expressive facial expressions of agent to achieve natural, harmonious and believable student-agent interaction. Based on the cues of sources and characteristics of facial expression, we propose a novel model of fuzzy facial expression generation, as seen in figure 6.

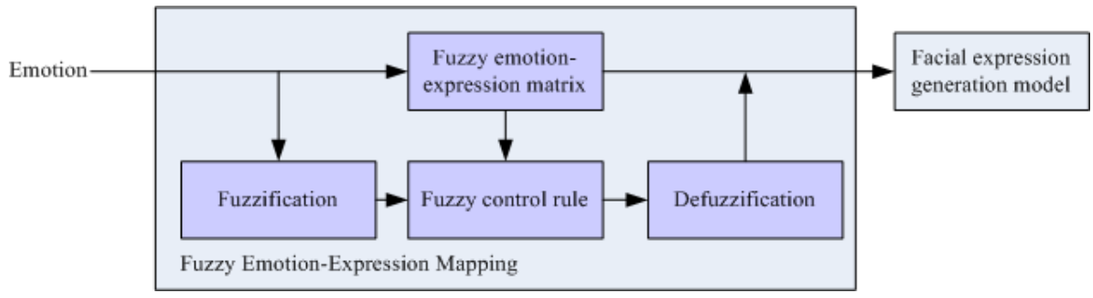


Fig. 6. Work flow of the textual affect sensing.

5.1.1 Fuzzy emotion-expression mapping

Fuzzy is one common characteristic of emotion and facial expression. There is also fuzzy relationship between emotion and facial expression. One emotion can be fuzzily expressed by multiple modes of facial expression. Here, we give the model of fuzzy emotion-expression mapping. The mapping of emotion to expression is one-to-many.

Based on the correlation of multiple facial expressions of emotion, fuzzy emotion-expression mapping is proposed, in which emotion and facial expression are supposed to be fuzzy vectors, and a fuzzy matrix consisting of degrees of membership maps the fuzzy motion vector to the fuzzy facial expression vector. Define the emotion space as $X = \{x_1, x_2 \cdots x_m\}$, where x_i is any emotion, such as surprise, disgust. Define the facial expression space as $Y = \{y_1, y_2 \cdots y_n\}$, where y_i indicates any mode of facial expression. The fuzzy relation $\tilde{R}$ from the emotion space X to the facial expression space Y is $R = (r_{ij})_{m \times n}$ where $r_{ij} = \tilde{R}(x_i, y_j) \in [0,1]$ indicates the correlation degree of (x_i, y_j) to $\tilde{R}$. Given the input emotional fuzzy vector E_S, the fuzzy facial

vector E_X can be obtained via fuzzy mapping, as seen in $E_X = E_S \circ R(ex_1, ex_2 \cdots, ex_n)$,where ex_i is the membership of the mode of facial expression y_i to the fuzzy facial expression $\tilde{E}_X$, $\circ$ means the compositional operation of the fuzzy relations. Once the fuzzy facial expression $\tilde{E}_X$ is determined, its intensity will also be computed. The intensity of selected emotion x_i is fuzzified to the linguistic value, which is then mapped to the linguistic value of related facial expressions according to fuzzy control rule. The intensity of facial expression y_i is obtained by defuzzifying its linguistic value. The emotion intensity and facial expression intensity also have fuzzy characteristics. The fuzzy linguistic values of emotion and facial expression are listed as very low, low, middle, high and very high. According to the emotion-expression intensity mapping, the mapping from linguistic value of emotion intensity to linguistic value of facial expression intensity was realized through fuzzy control. An example of fuzzy control rule is shown in Table 5. Emotion x (surprise) can be fuzzily expressed by facial expression y_1 or y_2. The very low intensity of x can be expressed by small intensity of y_1 or very small intensity of y_2.

Emotion x(surprise)	facial expression y_1	facial expression y_2
Very low	small	Very small
low	middle	small
middle	large	middle
high	Very large	large
Very high	— —	Very large

Table 5. Fuzzy control rule of fuzzy emotion-expression intensity mapping.

5.1.2 Facial expression generation model

The facial expression generation model is the module that accepts input of the fuzzy facial expression $\tilde{E}_X$ with its intensity and output the agent's facial expression. In this paper, we adopted Xface, an MPEG-4 based open source toolkit for 3D facial animation, to generate multiple facial expressions of emotions mentioned above. Figure 7 are some keyframes of facial expressions.

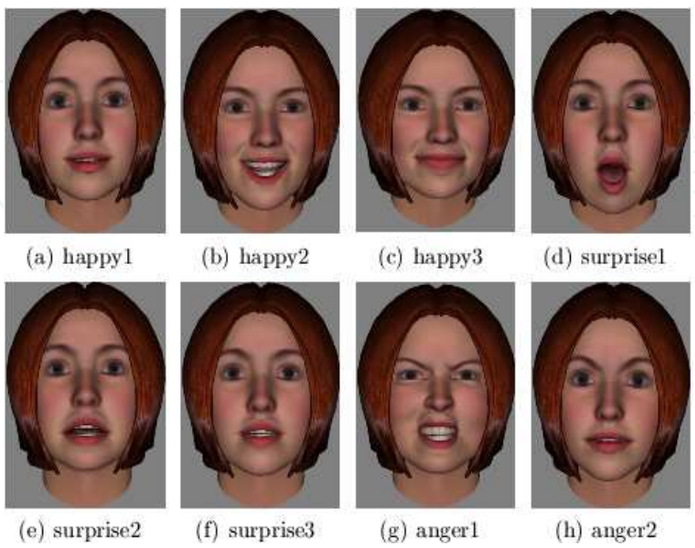


Fig. 7. Some keyframes of facial expressions.

5.2 Emotional speech synthesis

Speech is the easiest way to convey intention, and it is one of the fundamental methods of conveying emotion, on a par with facial expression. In this paper, the variety rule of prosodic features containing pitch frequency (F0), energy and velocity are concluded by analyzing emotional speech in our Emotional Speech Database. The autocorrelation function (ACF) method based on Linear Predictive Coding (LPC) and wavelet transform approach are employed to extract the F0 and tone respectively. Then prosodic features regulation is set up by utilization of Pitch Synchronous OverLap Add (PSOLA) and the original peace speeches are transformed into appointed emotional speech, including happy, anger, surprise and sad, based on the rules and regulation. Figure 8 illustrates the work flow of our approach.

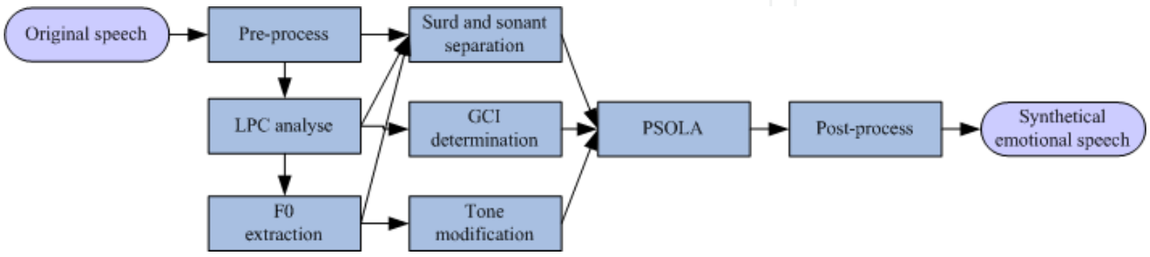


Fig. 8. Work flow of the emotional speech synthesis.

- **Pre-process.** Include noise elimination, pre-emphasis and amplitude normalization
- **LPC analyze.** Partition the original speech into frame, take LPC analysis of each frame, get the LPC residual function and first order reflection coefficient
- **F0 extraction.** Get the F0 through the autocorrelation analysis of the residual function and the F0 profile curve of the original speech
- **Surd and sonant separation.** Do the surd and sonant separation according to first order reflection coefficient, signal energy and frequency extraction result
- **Glottal Closure Instances (GCI).** determine the GCI according to the F0 extraction
- **Tone modification.** Extract the tone information using wavelet transform, modify the tone information according to the target-emotion; adopting inverse wavelet transform to get the F0 curve
- **PSOLA**. Using PSOLA technology to transform original speech into appointed emotional speech
- **Post-process.** De-emphasis, i.e. do the anti-operation of pre-emphasis in pre-process to restore speech effect

6. Tutoring strategies

Although our MITS can perfectly detect the attention information and affective state, it is also important to let the agent tutor know what to do with the information. As good human tutors can effectively adapt to the attention information and affective state of students, the most obvious way to learn about how to adapt the attention information and affective state of student is to learn from the human tutors. Sarrafzadeh videoed several tutors as they tutored students individually and a coding scheme was developed to extract data from each tutoring video to describe the behaviors, facial expressions and expression intensities of students and tutors. Tutoring actions are guided by a case-based method for adapting to student states that recommends a weighted set of tutor actions and expressions (Sarrafzadeh

& Alexander, 2008). However, this approach has two main shortcomings: firstly, the video only recorded three human tutors' behaviors and the students' reactions to these behaviors are not considered. What we want to get is the behaviors that can motivate the students, rather than arouse the students' averseness; and secondly, the coding scheme can not apply to the attention information and speech, text communication. In this paper, we use the traditional questionnaire to get the "optimal reaction" of the tutor towards the learner's attention information and affective state. The critical observation is that every excellent teacher has commonsense of the kind we want to give our agent tutor. If we can find good ways to extract commonsense from human tutor by prompting them, asking them questions, presenting them with lines of reasoning to confirm or repair, and so on, we may be able to accumulate many of the knowledge structures needed to give our agent tutor the capacity for commonsense reasoning for student's attention information and affective state. So we built a system called Human Tutor Commonsense make it easy for human tutors to collaborate to construct a database of commonsense knowledge. We invited more than 100 excellent teachers to log on our system to build the database. Then, a group of 50 students were asked to evaluate how much they satisfied with these commonsense, on a scale from 1 (strongly dissatisfied) to 5 (strongly satisfied). Then we chose two commonsense with highest mean score as the "optimal reaction" for each situation these questions described.
Based on the commonsense we obtained, MITS can be represented as a dynamic network as shown in Figure 9. Whenever the student's pedagogical state or attention information or affective state is changed, the following events are happen:

- Each time the dynamic network receives new evidence (the change of pedagogical state or attention information or affective state), a new time slice is added to the existing network
- The case-based method chose a tutorial action from the commonsense database
- The tutorial action is taken by the agent tutor "Alice"
- The history is updated
- "Alice" waits for the next student action

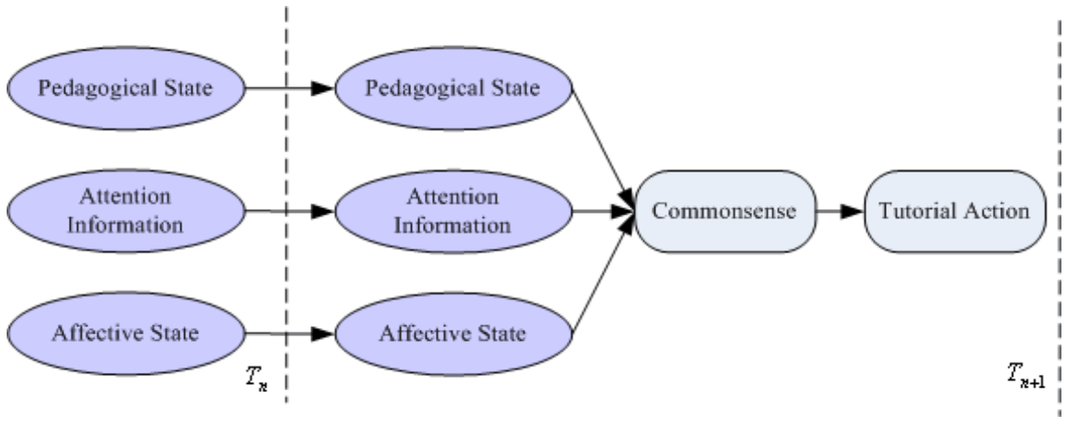


Fig. 9. Dynamic network for MITS.

7. Conclusion

This paper debuts a Multimodal Intelligent Tutoring Systems. Attention information detection and affective state detection are carried out. Meanwhile, the system adapts to the

student via an emotionally expressive agent tutor "Alice" through facial expression and synthetic emotional speech. Tutoring actions are guided by a case-based method that recommends a set of tutor actions and expressions for adapting to student states. The data that this case-based program uses were generated from questionnaires presented to human teachers.

In future work, it is necessary that the accuracy of emotion recognition and classification algorithm should be improved. Meanwhile, the MITS will be extended to integrate information from other sources including posture recognition and physiological channels such as pressure. We hope to evaluate the effectiveness of our system in a range of learning situations including more both young and adult learners. The test will provide more important directions for improvements to be made in the next version of MITS.

8. Acknowledgment

This work is supported by the National Nature Science Foundation of China (No.60873269) and International Cooperation between China and Japan (No.2010DFA11990).

9. References

Butz, J. PC., Hua, S., Maguire B.R., A web-based bayesian intelligent tutoring system for computer programming. Web Intelligence and Agent System, 4, 1(2006), 77-97

Chen, C.M., Intelligent web-based learning system with personalized learning path guidance, Computers & Education, 51, 2 (2008) ,787—814

DeSilva, L.C., Miyasato, T., Nakatsu, R., Facial emotion recognition using multimodal information, In Proc. ICICS1997 (1997), 397--401

Huang, T.S., Chen, L.C., Tao, H., Bimodal emotion recognition by man and machine, ATR Workshop on Virtual Communication Environments (1998)

Liu, H., Lieberman, H., Selker, T., A model of textual affect sensing using real-world knowledge. In Proc. IUI 2003, ACM Press (2003).

Mehrabian, A., Communication without words. Psychology Today, 2, 4(1968), 53-56

Qvarfordt, P., Zhai, S., Conversing with the user based on eye-gaze patterns. In Proc. CHI 2005, ACM Press (2005), 221-230.

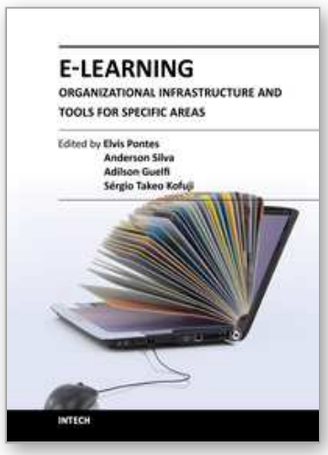
Reategui, E., Boff, E., Campbell, J.A., Personalization in an interactive learning environment through a virtual character, Computers & Education, 51, 2 (2008), 530-544

Roger, N., A framework for affective intelligent tutoring systems, In Proc. ITHET2006 (2006)

Sarrafzadeh, A., Alexander, S., Dadgostar, F., How do you know that I don't understand? A look at the future of intelligent tutoring systems. Computers in Human Behavior, 24, 4 (2008), 1342-1363

Siddappa, M., Manjunath, A.S., Knowledge representation using multilevel hierarchical model in intelligent tutoring system, In Proc. IASTED 2007 (2007), 323—329

Sun, P.C., Tsai, R.J., Finger, G., Chen, Y.Y., What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction, Computers & Education,50, 4,(2008), 1183-1202



E-Learning-Organizational Infrastructure and Tools for Specific Areas

Edited by Prof. Adilson Guelfi

ISBN 978-953-51-0053-9
Hard cover, 182 pages
Publisher InTech
Published online 17, February, 2012
Published in print edition February, 2012

Technology development, mainly for telecommunications and computer systems, was a key factor for the interactivity and, thus, for the expansion of e-learning. This book is divided into two parts, presenting some proposals to deal with e-learning challenges, opening up a way of learning about and discussing new methodologies to increase the interaction level of classes and implementing technical tools for helping students to make better use of e-learning resources. In the first part, the reader may find chapters mentioning the required infrastructure for e-learning models and processes, organizational practices, suggestions, implementation of methods for assessing results, and case studies focused on pedagogical aspects that can be applied generically in different environments. The second part is related to tools that can be adopted by users such as graphical tools for engineering, mobile phone networks, and techniques to build robots, among others. Moreover, part two includes some chapters dedicated specifically to e-learning areas like engineering and architecture.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen