ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 31

4th International Conference on Management Science and Management Innovation (MSMI 2017)

Online Learning Network Teaching System Based on Facial Recognition Technology

Xiangnan ZHAO
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
e-mail: 15120472@bjtu.edu.cn

Weidong ZHU
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
e-mail: wdzhu@bjtu.edu.cn

Bo WANG
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
e-mail: 14120474@bjtu.edu.cn

***Abstract*—Network teaching is a kind of teaching method for teachers to carry out teaching on Internet through the computer and Internet technology. Students wouldn't be limited by geographical and time, learning at anytime and anywhere and sharing high-quality teaching resources on internet. Network teaching is a truly effective modern education method. However, compared with the traditional classroom teaching, on network teaching, teachers can't distinguish whether students are interested in teaching content, whether students is concentrating on learning, and then make corresponding adjustment to teaching strategies by observing students' facial expressions. In order to solve this problem, this paper designed and implemented an online learning emotion detection system based on facial expression recognition technology. This system acquires students' online learning images through the webcam, analyses the focus of students' learning and their interest in teaching content through the three-dimensional learning emotion model we have established. This can help teachers to improve the content of teaching and comprehensively evaluate students' learning attitude, so as to promote the development of network teaching.**

Keywords-network teaching system; three-dimensional learning emotion model; learning emotion detection; emotion recognition

I. INTRODUCTION

Network teaching system is a form of an open learning environment which is based on Internet and information technology. With the advantages of real-time, convenience, and without the constraints of time and space, network teaching system combines all kinds of network services and teaching services to stimulate students' interest in learning and provides personalized tutoring. Compared with the traditional classroom teaching, network teaching can provide a convenient and rapid way for people to learn because of the characteristic of space-time separation, but there still has some drawbacks. In network teaching, teachers and students can't communication face to face, so that teachers can't analyze students' learning emotion and the understanding level towards to the learning content by observing his facial expressions, and then adjust teaching strategies according to the condition of students' learning. If things go on like this, students can't get attention of teachers in the learning process, learning problems can't be resolved in a timely manner. These will cause a decline in interest in learning, and even affect physical and mental health of students. So the development of network education requires network teaching system has the ability of detecting emotion, teachers can make an accurate tracking on students' learning emotion, and make different teaching strategies according to the different emotions in learning process of students [1].

Aiming at the emotional defects of network teaching system, we designed and implemented the emotion detection network teaching system based on facial expression recognition in this paper. Compared with the traditional network teaching system, this system not only can identify and analysis the learning emotions of learners, but also converts video and text information on client side, which greatly reduces the storage of learning platform server and network load pressure. Hence, teachers can know the learning state of students without watching learn video monitoring.

II. ONLINE LEARNING EMOTION MODELING

Based on the basis study of learning state space by Wang Zhiliang and others, this paper defines the three-dimensional learning emotion model, as shown in Fig. 1. The X axis in Fig. 1 represents the cognition degree of students towards to the learning contents, the Y axis represents the aversion degree and Z axis represents the excitement degree. So we can establish a three-dimensional learning emotion model to analyze the expressions characteristics of students in the different learning process.

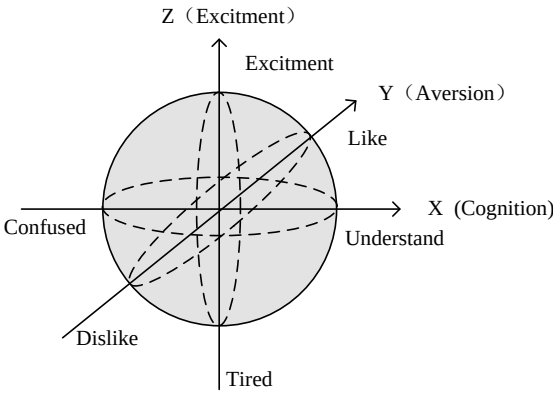

Figure 1. Three-dimensional learning emotion model

1) Cognition

Cognition refers to students' cognition degree towards to teachers' teaching content. Students will express happy mood when he is understanding the learning content, facial expressions will appear as the cheeks muscles lift and mouth upturned, etc. When students can't understand the learning content, the confused emotion will show up, at the same time the face will also appear pouting, cheeks stretch, eyebrows turn into an eight character, etc.

2) Aversion

Aversion refers to students' interest level to teachers' teaching content. When students enjoy the learning content, they will open up their eyes and stare at screen all the time. When students dislike what they are learning or the teachers' teaching method, the aversion feeling will show up. The face will appear the down curved mouth and eyebrows.

3) Excitement

Excitement refers to students' mental state. Students will become exciting when they have a good mental state in the process of learning and the face will appear corresponding expression, such as eyes sparkling, mouth slightly rising, etc. However, the students will be tired for a long time continuous study, the face will appear frequently blinking, yawning and even sleep expression.

Expression is the reflection of students' inner world, and the facial expressions of students reflect their learning emotions and state, so teachers need to pay attention to the facial expressions of students in teaching. Through the study of the expression in students' learning process, this paper established a three-dimensional learning emotion model, which provides a reliable mathematical model for the study of emotion detection in network teaching.

III. LEARNING EMOTION RECOGNITION

A. Face Feature Tracking Based on CLM Algorithm

In the fields of face feature tracking, the Active Shape Models (ASM) [2] was first proposed by Tim et al. in 1995, which is the most widely used face algorithm using in feature point location. ASM uses constraint relationship among the facial feature points and restricts the shape parameter through the training model, to make the searched feature points under a reasonable shape constraints, which can be more accurate to locate the feature points.

Tim et al. made further improvements to the ASM algorithm and proposed the Active Appearance Model (ASM) [3]. ASM is operated on the basis statistical shape model [4], while based on ASM, AAM further makes texture features modeling and combines the shape model and texture model as an appearance model. It includes not only the features of non-rigid shape transformation, but also the texture information of all feature points around. AAM improves the accuracy of ASM model by using global texture information, but it also has the following two shortcomings: First, using global texture information makes the model has a high dimension, as a result AAM has a poor real-time; Second, using gray value of image as the texture feature directly is too simple, which results in an inefficient AAM algorithm.

Constrained Local Model (CLM) was proposed based on the idea of AAM and ASM [5], it not only inherits the advantages of ASM and AAM algorithm, but also improves above two disadvantages, which make great breakthrough in accuracy and real-time. CLM also includes two models, the shape feature model--the constraint model, and the texture feature model around feature points--the local model.

CLM has the same shape model as ASM and AAM, but its texture model is different. CLM takes each feature point as the center, then selects a neighbor block and normalize it, after then connects all neighbor blocks into a gray vector, which is used as the texture vector of image. For the local block of each feature point, there needs to train a linear Support Vector Machine(SVM) to identify it, which can be used in the subsequent search. The fitting formula of shape model and texture model is shown in (1).

$$\begin{aligned} x&=\bar{x}+P_s b_s \\ g&=\bar{g}+P_g b_g \end{aligned} \tag{1}$$

Where $\bar{\boldsymbol{x}}$ means the average shape, P_s means the principal component matrix of shape model, b_s means the non-rigid variation of the average shape of shape model. Analogously, $\bar{g}$ represents the average normalized strength vector of local texture feature model, P_g means the principal component matrix of the local texture feature model and b_g means the non-rigid variation of the average shape of local texture feature model.

After CLM model is constructed, the face shape model can be initialized at any position of the training sample image. Aiming at the shortcoming of ASM algorithm, the main idea of the following improvement is to model the neighborhood of feature points. Since the CLM algorithm is based on the local SVM response graph. Using SVM classifier to train the texture around each feature point and using face model to initialize face, which can be applied to the subsequent search.

CML algorithm converges quickly, so it can be used in video face tracking and act on each frame in the video image. However, in actual video images, the position, pose, size and other information of two adjacent video images will not change too much. While, in the process of CLM searching,

each video frame needs to be initialized by the average model which is trained, thus each frame wastes more time to search. Therefore, the CLM model of the previous frame can be used as the initial position of next video image, so it can converge to the position of current video frame more quickly.

By using previous frame position to optimize current frame position, the time of face search is greatly reduced. In order to obtain an intuitive time comparison results, we used two methods to analyze a short video. In this paper, we intercepted 20 frames of them to compare their performance. As shown in Fig. 2, when using the average model on each frame the processing time is about 26ms. In contrast, using information of previous frame can significantly improve the speed of face search, it costs about 10ms.

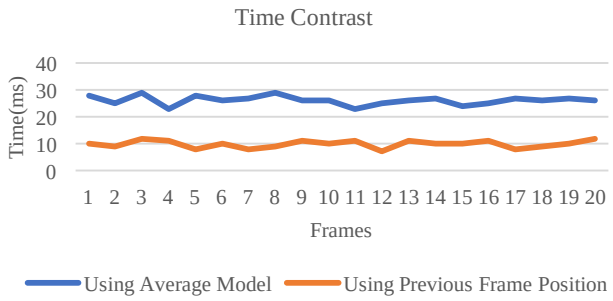

Figure 2. Time comparison of video detection optimization

B. *Learning Expressions Classification Based on SVM*

1) *Basic principles of SVM*

Support vector machine (SVM) [6] is a supervised learning model for analyzing data in classification and regression analysis. The classification principle of SVM can be summarized as finding a hyperplane in an n-dimensional space to make the instances of training set separated in n-dimensional space by this hyperplane, and the distance between the classified instances point and the hyperplane as large as possible.

As for learning emotions, each expression of learners can be regarded as the synthesis of base vectors of three-dimensional learning emotion model. Therefore, there is no need to classify the learning expressions precisely, instead of it, we divided it into several basic learning expressions, so using $y \in \{-1,+1\}$ to categorize the information may be not appropriate. For this multi-class problem, SVM breaks down the original problem into some sub problems, so as to obtain the problems which can be classified directly, then combines the classification results of the sub problems according to some certain rules, thus we can get the results of the original problem. Here we used the one-to-one classification algorithm in this paper. The core idea of this algorithm is to train a binary classifier for each two categories. That is, for a k class problems, there need to train $k(k-1)/2$ binary classifiers. When classifying the instances, the input instances are successively classified by $k(k-1)/2$ binary classifiers, then we will get $k(k-1)/2$ classification results. Finally, we combine all the results by using a voting method, is concrete is, each binary classifier vote for its classified side, then count the number of votes, the most votes is the final results of multi-class classifier. In this paper, we divided learning expressions into six basic expressions through the three-dimensional learning emotion model.

Posterior probability can be used as a practical pattern classification model. For a k class problems, any one input X, the target is to estimate $p_i=P(i)=P(y=i|X)$, $i=1, 2, \cdots, K$. Using one-to-one method, SVM trains a binary classifier for each two categories, and the probability is r_{ij}. The formula is show in (2):

$$r_{ij} \approx p(y=i|y=i \text{ or } j, X) = 1/1+exp(Af+B) \quad (2)$$

Then we use the method of voting to get the comprehensive probability that is $P(i)$, as it is shown in (3).

$$P(i)=\frac{2}{k(k-1)}\sum_{j:j\neq i} I_{\{r_{ij}>r_{ji}\}}$$
$$I(x)=\begin{cases}1 & \text{if } x \text{ is true}\\ 0 & \text{if } x \text{ is false}\end{cases} \quad (3)$$

2) *Algorithm training and testing*

We collected 244 facial expression images as the expression database for algorithm training, including 40 images for understanding expression, 42 images for confused expression, 39 for like, 40 for dislike, 40 for exciting, and 43 for tired. We used 124 images of the samples for algorithm training and another 120 images for algorithm testing. The kernel function used by SVM is $k(x, y)=exp\left(-\gamma \left\|x-y\right\|^2\right), \gamma>0$, of which $\gamma=0.9$, and $b=1$ in decision functions.

The algorithm testing included two parts, the first is to test the one to one method for classifiers. 20 test samples were used for each learning expression, the test results are shown in Tab. 1. The total correct rate is 93.3%, in which 2 images of understanding are identified as like, 1 images of confused is identified as tired, 2 images of like are identified as exciting, 2 images of dislike are identified as confused and 2 images of tired are identified as confused.

TABLE I. EXPRESSIONS TEST RESULTS

Expression	Samples	Results	Correct	Rate
Understand	20	18	18	90%
Confused	20	23	19	95%
Like	20	21	19	95%
Dislike	20	18	18	90%
Exciting	20	21	20	100%
Tired	20	19	18	90%

Another part of algorithm test is the probability output on the basis of one-to-one method. As we can see form Fig. 3, there are the test results of two cognitive test samples. The first image is identified as understanding at the probability of 0.6, the probability of dislike is 0.2 and the

sum of other probability is 0.2. As for the second image, the probability of confusion is 0.6 and the probability of dislike is about 0.4.

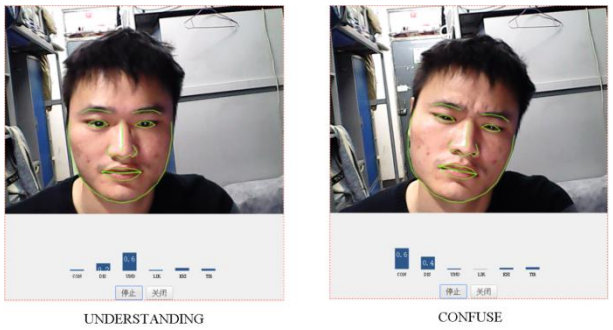

Figure 3. Cognitive learning expression recognition results

As shown in Fig. 4, there are the test results of two aversion test samples. The first image is identified as like at the probability of 0.4, the probability of understanding is 0.3, the probability of exciting is 0.2 and the sum of other probability is 0.1. As for the second image, the probability of dislike is 0.6, the probability of understanding is 0.3 and the sum of other probability is 0.1.

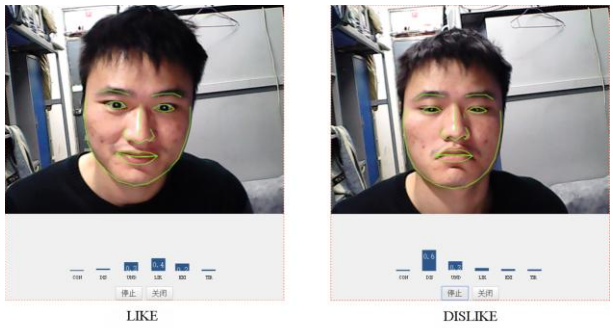

Figure 4. Aversion learning expression recognition results

As we can see form Fig. 5, there are the test results of two excitement test samples. The first image is identified as exciting at the probability of 0.9 and the sum of other probability is 0.1. As for the second image, the probability of tired is 0.6, the probability of exciting is 0.2 and the probability of like is 0.1.

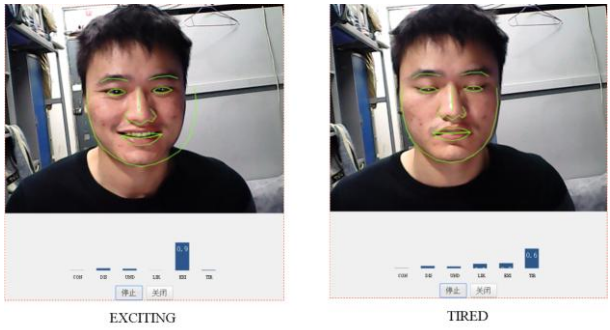

Figure 5. Excitement learning expression recognition results

IV. Online Learning Emotion Detection Network Teaching System

A. System General Introduction

This system is mainly used for online learning monitoring and learners' emotion analysis. It has realized the function of video image acquisition, image preprocessing (including image enhancement, gray image, LBP feature extraction and normalization), face recognition, expression feature extraction and classification, learning emotion analysis. We selected Java to take the main development language, MySQL to take the database, HTML5 to take video capture and achieved face recognition and expression recognition by JavaScript. In order to improve communication security, Web browser and server use HTTPS communication method. The entire development process is carried out under Ubuntu 16.04 LTS operating system and the final project running on Windows Server 2008.

Compared with the traditional network teaching system, this system is not only compatible with a variety of electronic equipment, but also feedbacks the learning status of students to teachers directly, which can greatly reduce the pressure on learning platform server storage and network load. After teachers assign the learning task, students can learn in any place. Due to HTML5 is compatible for each device, learning terminal can be computer, notebook, PAD, mobile phone and other electronic products with a camera. In this system, face recognition and expression recognition are realized by JavaScript, so there is no need to upload the video information to server, instead, we get the recognition result on client, and send it to server in text information form.

B. Learning Emotion Analysis

We label the classification results of learning expressions obtained from the probability based SVM classifier as: $x_{understand}$,$x_{puzzled}$,y_{enjoy},$y_{disgust}$,$z_{exciting}$,z_{tired}. So, the result of classification can be converted into coordinates of three-dimensional learning emotion model by using the (4).

$$\begin{aligned} x&=x_{understand}-x_{puzzled} \\ y&=y_{enjoy}-y_{disgust} \\ z&=z_{exciting}-z_{tired} \end{aligned} \tag{4}$$

Where $-1 \le x, y, z \le 1$. Since the process of acquisition and recognition are all realized by the client, the number of video images processed every second by terminals of different configurations can be different. On the other hand, due to the continuity of learning expressions. So, assuming that the processing rate is vFPS, we can use the (5) to normalization of data per second.

$$\bar{\chi}=\frac{1}{v}\sum_{i=1}^{v}\chi_i \tag{5}$$

In Eq. (5), $\chi \in$ { understanding, confused, like, dislike, exciting, tired}. Through the normalization, on the one hand, we can reduce the influence of dirty data, on the other hand, the effect of different configuration terminals on the data can be eliminate.

We collected and analyzed the learning data of 58 students who have studied online for a period of time in a course, with the distance between students and cameras is about 50cm. Experiment shows that the learning expressions data of the top 15 students is more than 50% located in the first dimension of the three-dimensional coordinates. So we can draw the conclusion that the higher proportion of learning data in first dimension, the better learning effect can be got.

TABLE II. STUDENT LEARNING EMOTION DATA DISTRIBUTION

Dimension	Points(A)	Points(B)	Percent(A)	Percent(B)
1	2558	537	71.0%	22.8%
2	183	128	5.1%	5.4%
3	79	105	2.2%	4.5%
4	180	661	5.0%	28.1%
5	108	218	3.0%	9.3%
6	78	264	2.2%	11.2%
7	184	228	5.1%	9.7%
8	233	213	6.5%	9.0%
SUM	3603	2354	100%	100%

The learning data of two students is shown in Tab. 2, The learning time of student A is 1 hours, we collected 3603 learning expressions, the student B learned for 40 minutes and 2354 learning expressions was gained. From Tab. 2, we can see that the learning expressions of student A take 71% in first dimensional, that is to say, the classification result of SVM classifier is located at the dimension of understanding-like-exciting takes 71% proportion, while student B takes only 22.8%. It shows that compared with student B, student A have a better serious learning attitude and a better learning state. In fact, the student A focused on learning and got score of 87, while the student B often had his eyes closed and yawned, so the final score of student B is 62, so in line with the above test results.

Through the analysis of the learning expressions data collected in this system, there will be two main influence on teaching: First, teachers can know students' learning state through the analysis results of the system. Second, through analyzing the learning emotions of students, teachers can get the degree of students' interest in course.

V. CONCLUSION

Aiming at the emotion absence in network teaching, based on three-dimensional learning emotion model, this paper designed and implemented the online learning emotion detection network teaching system by using face detection technology, face tracking technology and facial expression recognition technology. This system realizes the personalized network teaching on cognition and emotion levels. This is the development trend of network teaching system. At the end of this paper, we carried out the experiment, and verified the feasibility of this system through the analysis of the learning data of students in a course.

REFERENCES

[1] Ren Y. The Design and Implementation of Network Teaching Platform Basing on. NET[J]. Physics Procedia, 2012, 25:892-898.

[2] Cootes T F, Taylor C J. Data Driven Refinement of Active Shape Model Search[M]. 1996.

[3] Cootes T F, Hill A, Taylor C J, Haslam J. The Use of Active Shape Models for Locating Structures in Medical Images[J]. Image & Vision Computing, 1993, 12(6):355-365.

[4] Heap T, Hogg D. Extending the Point Distribution Model using polar coordinates[J]. Image & Vision Computing, 1995, 14(8):589-599.

[5] Saragih J M, Lucey S, Cohn J F. Deformable Model Fitting by Regularized Landmark Mean-Shift[J]. International Journal of Computer Vision, 2011, 91(2):200-215.

[6] Osuna E E. Support vector machines: Training and applications[C]// 1970:1308-16.