Inferring Academic Emotion in Online Learning based on Spontaneous Facial Expression

Cun-Ling BIAN[a], De-Liang WANG[a], Ya ZHANG[a], Wei-Gang LU[a,b*]

[a] *Department of Educational Technology, Ocean University of China , China*

[b] *Department of Computer Science and Technology, Ocean University of China, China*

* luweigang@ouc.edu.cn

Abstract: Academic emotion is one of the important factors impacting on learning effect. A robust automatic academic emotion inference method will have significant meaning for the educational field. This paper presents a study on the relationships between spontaneous facial expressions and academic emotions when students take online learning courses. First we establish corpora. The images from 82 students are collected with high definition cameras in an almost natural environment. Both students and external coders are invited to label the corpora. Then, the methodologies of academic inference algorithms are described based on both artificial feature and Convolution Natural Network (CNN). A preliminary experiment inferring self-annotation academic emotion is carried out to validate their actual effect. Among all the algorithms, the CNN-based algorithm using some tricks exhibited the ability to infer learners’ academic emotion from the learner’s expression has the highest accuracy. This study has potential value to make up for emotion absence existing in online learning.

Keywords: Academic Emotion Inference, Spontaneous Facial Expression, Machine Learning, CNN

1. Introduction

Emotion is one of the important factors impacting on learning effect (Newton, 2013). The learning-related emotion is also called academic emotion. Although it is an inner psychological reaction, a seasoned teacher can still grasp the emotion through exterior behaviors from students. As the most visible reflexes of emotions, facial expression is an important study objective (Ekman, 1984). One hot study topic is inferring emotions manifested by facial expressions with the help of a computer. It is a complex task, which involves the study of computer version, machine learning and behavioral science (Lucey, Cohn, Kanade, & Saragih, 2010). However, it also has some significant advantages compared to other inference methods. For example, facial expression recognition (FER) is non-contacted, which implements the function with a minimal interference. FER has been one of the most common technologies in educational affective computing field (Monkaresi, Bosch, Calvo, & D’Mello, 2017).

Collecting the data, obtaining the ground truth, and defining the appropriate emotion model are some of the challenges in building any affective model (Picard, 2003). And a few of publicly available facial expression datasets relate to academic emotions now. Therefore, an online-learning spontaneous facial expression corpora is established in this paper. Basically, five meaningful academic emotions are included, respectively, confusion, neutral, distraction, fatigue, and enjoyment. The labels are marked by both participants and external coders, which are also validated by Fleiss’ kappa (Fleiss, 1971) and Cohen’s kappa. With these high quality corpora, it is hopeful to identify the effect of some inferring algorithms in the next stage for the purpose of getting more insights of this study’s application value.

2. Methodology

2.1 Corpora

Academic emotion belongs to the subtle affective state, which cannot be described using the basic emotions (Pekrun, Goetz, Titz, & Perry, 2002). However, there is still no consolidated definition of it. Hara and Kling report that frustration, isolation, anxiety, and confusion are the most frequent feelings that students experienced in online learning (Noriko, & Kling, 2003). But, You et al. identify enjoyment, confidence, fear, frustration, boredom, and anxiety toward the learning system as conventional emotions in online learning (You, Kim, & Park, 2012). In these corpora, the academic emotions included must have relatively apparent facial expression features. Therefore, five meaningful academic emotions are selected, namely enjoyment, confusion, fatigue, distraction, and neutral. The encodings are defined as follows: the academic emotion of confusion is coded if the student seems puzzled and not sure how to continue or is struggling to understand the material; enjoyment appears when the student shows interest or feels enjoyable in current course; fatigue will descend on the student when he feels drained of strength and energy or looks lethargic and yawns; distraction is coded if the student gets his mind off the screen combining the concrete circumstance of online learning. The last academic emotion, neutral, is coded when the student does not express any mood.

82 healthy students between the age of 17 and 26 (mean age = 20.09; standard deviation = 2.26; male = 29; female = 53) voluntarily participate in the experiment. They all sign informed consents and have the right to quit the experiment at any time. The corpora only include the images of the participants who sign the consent form agreeing to their facial images being used for study purposes. Before the experiment, the participants' characteristics including learning styles, content preference and cognitive level are identified through investigation and statistics. Then some personalized online video courses collected from MOOC platforms are prepared to elicit different academic emotions (Jaggars, & Xu, 2016). The durations of these stimuli range from about 10 minutes to roughly 60 minutes. Among the courses, some with appealing content and humorous teaching styles are chosen to elicit enjoyment; some appealing but high-level cognitive mathematical and logic courses are selected to elicit confusion; some unattractive and long-winded courses are prepared to elicit fatigue and distraction academic emotions; the neutral are collected during the entire process. Before the formalized experiment, the effectiveness of these stimuli has been verified by a pretest.

The whole experiment is conducted in a study hall, where a normally functioning computer is prepared to play the online courses. A webcam is placed above the screen, which records video at 30 frames per second with a resolution of 1280 × 720. The recordings throughout the learning process will be saved in the experiment. Besides, the illumination and occlusion conditions are not strictly controlled here. Actually, the differences between the experiment and the real world are minimized to reduce Hawthorne effect that the experiment will affect the natural behavior of participants as well as guarantee the objectivity and authenticity of data. The subjects can participate in the experiment anytime during the day or night and the fluorescent lights are on when the light is poor in the room. Therefore, the brightness changes in a large range. There also exist some obstacles which disguise important facial expression information, such as spectacles, hair, and hands. Although these confounding variables increase the hardness of recognition, they can make the corpora more realistic and are beneficial for evaluating the robustness of algorithms. In one sentence, the facial expressions captured in an almost natural environment.

It has been emphasized that participants could not know the experiment's true intention when getting spontaneous facial expression (Sebe, Lew, Sun, Cohen, Gevers, & Huang, 2007). Therefore, the participants are directed to the study hall and informed to take some online lessons for curriculum assessment. Then, the stimuli are displayed in the order of enjoyment, confusion, distraction and fatigue. Afterwards, the participants are told the experiment's true intention and asked to watch their own video recordings to finish a subjective emotion assessment form, since participants themselves often provide a useful source of information about their emotions (Porayska-Pomsta, Mavrikis, D'Mello, Conati, & Baker, 2013). If the participants find a facial expression belongs to a specific academic emotion, they should record the academic emotion label, the starting time and ending time of this expression. Except for the five pre-defined academic emotions label, the participants can also add others if they want. One advantage of the retrospective reporting is that it eliminates the problem of increased cognitive load and interference with the learning activity. In addition to that, the retrospective reporting can also offer to the participants a possibility to verbalize and discuss those reflections with a researcher, which may contribute to the students' development of better emotion judgment (Porayska-Pomsta, Mavrikis, D'Mello, Conati, & Baker, 2013). Before the self-annotation, all the participants are trained to select video clips and make self-annotations for academic emotion to control the quality. A total of 1274 video clips are segmented with self-annotation from the recordings. Besides, four coders are also invited to make external annotation on the segmented video clips. The external annotation has fairly high confidence that all reports of an emotion state involve the same constructs and unlikely to have intentional bias (Porayska-Pomsta, Mavrikis, D'Mello, Conati, & Baker, 2013). Each video clip is converted to an image sequence after face detection and composes the image corpora.

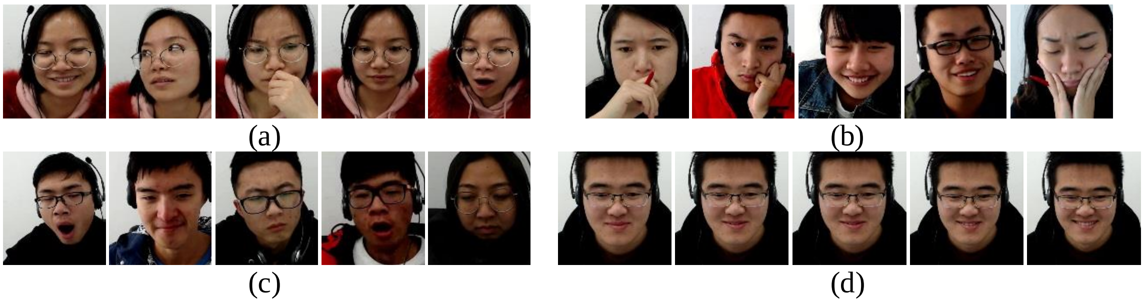

Figure 1. (a) All five academic emotions with the same participant, (b) Some expression images of different occlusions, (c) Some expression images of different illuminations, (d) The different intensity of same expression.

The final corpora include 30,124 facial expression images. The number of confusion, neutral, distraction, fatigue and enjoyment are 5292, 8278, 4866, 5524, 6224, respectively. All five academic emotions of one participant are displayed in Figure 1(a), which are enjoyment, distraction, confusion, neutral, and fatigue from left to right. However, not every participant shows all five academic emotions in the corpora. Since the spontaneous expression of the same academic emotion an individual shows varies with the situation and time, several expressions of the same participants are included in the corpora. Some expression images with different occlusions and illuminations are demonstrated in Figure 1(b) and 1(c), respectively. In the corpora, 62.2% of the participants are bespectacled, and some facial expression features are covered by hands, hair. The intensity of an expression varies in different phases because each clip records the facial expression from onset to apex. Figure 1(d) illustrates the variation. The Fleiss's kappa among external coders is 0.867 while the Cohen's kappa between the external and the self-annotation is 0.943, where the external label is decided by the majority of the trained coders. It is credible that the annotations in our corpora are high-quality. The corpora is made publicly accessible, which will certainly be useful in educational affective computing, especially the practicality of academic emotion inference. If you want to get the corpora, the access can be requested by emailing the corresponding author.

2.2 Academic Emotion Inference Algorithms

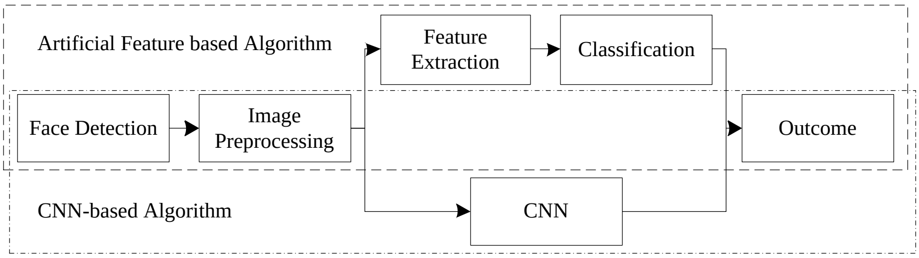

Figure 2. The flow diagram of inference algorihms.

To realize automatically academic emotion inference, two kinds of inference algorithms have been implemented. The flow diagram is shown in Figure 2. First they need to detect and crop the face region from the input image and do preprocessing operations like illumination normalization. Then, the artificial feature based algorithm has to take feature extraction and classification while the CNN-based algorithm only relies on CNN. Specifically, Local Binary Pattern (LBP) is taken as the artificial feature, and K-nearest neighbors (KNN) or Support Vector Machine (SVM) is the classifier in this study. LBP is a particular case of Texture Spectrum model, which compresses the pixel value in a location to its neighboring pixel value and generates a binary number representing the pattern. The feature of an image is usually extracted as the histogram of LBP to cope with illumination and rotation variation. In this study, the face image is divided into 5×5 non-overlapping sub regions at first. Then, the histogram of each region is concatenated as a vector, so the detail and position of the pattern are captured. KNN, an adopted classifier, is a type of instance-based learning, or lazy learning, whose function is only approximately local and all computations are deferred until classification. SVM is a supervised learning algorithm, which intends to create an optimal separating hyper plane between different classes to minimize the generalization error and thereby maximize the margin. The SVM based on radial basis function kernel is used here.

CNN is an end end-to-end algorithm, where the feature extraction and classification are integrated (Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Dumitru, Vanhoucke, & Rabinovich, 2015). It is becoming more and more popular in the computer vision field. The framework of CNN usually consists of several convolutional layers and fully connected layers. And between them, there may also be pooling layers, normalization layers and other special layers to improve the performance. In our study, VGG16 is adopted as the concrete CNN architecture, which has been proved to be a good choice in some computer version task (Ren, He, Girshick, & Sun, 2017). The VGG16's input is an image of fixed-size 224 × 224 × 3. Most of the convolutional layers in VGG16 use a convolution kernel 3 × 3 with a small receptive field. However, some convolutional layers utilize 1 × 1 kernel which can be seen as a linear transformation of the input channels. To preserve the spatial resolution, all convolutional layers take 1 pixel stride in convolution operation. One point needs to be emphasized: the output size of the last full connected layer in VGG16 is adjusted to five which represents the numbers of academic emotions

2.3 Training Models

The experiment using facial expression images to infer the self-annotation academic emotion is conducted on the above corpora. Numerically, the corpora are enough to train both KNN and SVM. However, because CNN models often have too many undetermined parameters, training them need a large number of samples as the basis. Without sufficient samples, it will lead to the over-fitting. Therefore, some necessary tricks must be adopted in our study to counter the limited corpora. The first one is transfer learning where the basic idea is to utilize the feature extraction module and adapt the classification module trained in the source-domain to the target-domain (Shin, Roth, Gao, Lu, Xu, Nogues, Yao, Mollura, & Summers, 2016). In detail, the pre-trained VGG16 model is selected to initialize the network parameters and only a part are adapted relying on training data. Because the CNN model makes no sense without transfer learning, the trick is adopted by default. Data augmentation is another effective trick, which is used to expand the existing data set. It can be implemented in many ways, such as adding noise and applying affine transformations (Han, Liu, & Fan, 2018). The ways include scaling, rotating, translating, dropout, additive Gaussian noise, element wise multiply, perspective transforming, cropping and padding. Each original image is randomly changed twice, based on one of the nine forms each time. Each output is saved as a new sample. Therefore, the size of the corpora is doubled.

It must be stated that training CNN-based model consumes massive computing resources while artificial feature based model is relatively small. In our study, a workstation with an Intel Xeon E5-1650 V4 CPU and 32GB RAM is prepared as the experimental platform. However, two NVIDIA GeForce GTX 1080Ti GPUs are used to complete the training process of CNN model within a limited time.

3. Preliminary Results

The performances of different algorithms are presented in Table 1, where the precision, recall, F1, average accuracy and kappa score for each algorithm are provided by using ten-fold cross validation over participants. The recall and the precision are intuitively the abilities of the algorithm to find all the positive samples and not to label a negative sample as positive, respectively. The accuracy is the ratio of samples that have been correctly classified. The F1 score can be interpreted as a weighted average of the precision and recall. The kappa is a statistic that measures the internal consistency between predictions and real labels. CNN-DA is the CNN-based algorithm that adopts data augmentation. The results show that although CNN is powerful, the CNN-based algorithms are not always the best. LBP-SVM is superior to both LBP-KNN and CNN in all indictors with 0.751 accuracy and 0.684 kappas. However, after adopting data augmentation, the CNN-based algorithm is significantly improved, 0.877 accuracy with 0.880 F1, which is also the state-of-the-art algorithm. Due to the limited sample size and subtle expression in these corpora, the algorithms are underperforming in it compared with others. But overall,

The preliminary results indicate that inferring academic emotion based on facial expression images is feasible.

Table 1: The results for different algorithms

Method	Precision	Recall	F1	Kappa	Accuracy
LBP-KNN	0.361	0.356	0.357	0.142	0.328
LBP-SVM	0.744	0.762	0.747	0.684	0.751
CNN	0.641	0.599	0.563	0.450	0.560
CNN-DA	0.908	0.861	0.880	0.840	0.877

4. Conclusions and Future Work

In this paper, we describe a methodology for inferring academic emotion based on spontaneous facial expressions from students in online learning. In the process, a spontaneous facial expression corpora is created. Then, the experiment conducted on the corpora demonstrates the effectiveness and reliability of our methodology. The preliminary results show that our methodology is feasible and CNN-based algorithm using some tricks is superior. However, we only use the facial expression images from learners in this experiment. The future study will consider various clues, such as behaviors and speeches, to improve the academic emotion inference.

Acknowledgment

This work has been supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 16YJC880057) and the Fundamental Research Funds for the Central Universities (No. 201564041).

References

Ekman, P. (1984). Expression and the nature of emotion. *Approaches to Emotion*, 3:19–344.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Han, D., Liu, Q., & Fan, W. (2018). A new image classification method using CNN transfer learning and web data augmentation. *Expert Systems with Applications*, 95:43–56.

Hara, N., & Kling, R. (2003). Students' frustrations with a Web-based Distance Education Course: An Ethnographic Study of Participants' Experiences. *Turkish Online Journal of Distance Education*, 4(2):69–70.

Jaggars, S. S., & Xu, D. (2016). How do online course design features influence student performance?. *Computers & Education*, 95:270–284.

Lucey, P., Cohn, J. F., Kanade, T., & Saragih, J. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *Computer Vision and Pattern Recognition Workshops*. IEEE, 94–101.

Monkaresi, H., Bosch, N., Calvo, R., & D’Mello, S. (2017). Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*, 8(1):15–28.

Newton, D. P. (2013). Moods, emotions and creative thinking: A framework for teaching. *Thinking Skills & Creativity*, 8(1):34–44.

Pekrun, R., Goetz, T., Titz, W., & Perry, RP. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37(2):91–105.

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies*, 59(1):55–64.

Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., & Baker, RSJD. (2013). Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education*, 22(3): 107–140.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Sebe, N., Lew, M. S., Sun, Y., Cohen, I., Gevers, T., & Huang, T. S. (2007). Authentic facial expression analysis. *Image and Vision Computing*, 25(12): 1856–1863.

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5): 1285–1298.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Dumitru, E., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Computer Vision and Pattern Recognition*. IEEE, 1–9.

You, J., Kim, H., & Park, S. H. (2012). Development and construct validation of e-learning academic emotion scale (e-AES). *The Journal of Yeolin Education*, 20(3): 19–44.