EPiC Series in Computing
Volume 69, 2020, Pages 224–229
Proceedings of 35th International Conference on Computers and Their Applications

Estimating the Concentration of Students from Time Series Images

Nguyen Hoang Son[1], Yu Takahata[1], Masaaki Goto[1], Tetsuo Tanaka[1], Akihiko Ohsuga[2] and Kazunori Matsumoto[1]
[1] Graduate School of Engineering, Kanagawa Institute of Technology, Japan
{matumoto, t-tanaka}ic.kanagawa-it.ac.jp
[2] Department of Informatics, The University of Electro-Communications, Japan

Abstract

In this study, we build a system that is able to estimate the concentration degree of students while they are working with computers. The purpose of learning is to gain knowledge of a subject and to reach sufficient performance level about the subject. Concentration is the key in the successful learning process. But the concept of concentration includes some ambiguity and lacks the clear definition form an engineering point of view, and it is difficult to measure its degree by observation from outside. We in this paper begins with a discussion of the concept of concentration, and then a discussion of how to measure it by using standard devices and sensors. The proposed system investigates the facial images of students recorded by the PC webcams attached to the computers to infer their concentration degree. In this study, we define the concentration degree over a short time interval. The value takes continues value from 0 to 1, and is determined based on the efficiency of simple work performed over the interval. We convert the continuous values into three discrete values: low, middle and high. In the first approach in this study, we apply deep learning algorithm with only the facial images. In the next, we obtain the data of face moves as a set of time series, and run the learning algorithm using both of the data. We explain an outline of the methods and the system with several experimental results.

1 Introduction

The final achievement performance level is important in learning. Many education technologies have been developed to reach higher levels of performance about a subject in a shorter time with less efforts. The motivation and willingness of a student to learn, and his/her interest in the content are the influential factors to the learning efficiency [9, 12]. High motivation and interest lead to high learning efficiency, and lead to the learning success. However, these factors cannot be observed directly because they are the mental state of the student. On the other hand, the degree of concentration appears through various types of behaviors of students, such as facial expressions, moves of the body parts, voices and so on. When the student is willing to work on an interesting task, he/she is focused with high concentration, quiet and calm. Skilled teachers carefully observe these behaviors in the classroom to guess students' interests and motivations.

This research aims to realize this ability of human on the computer by using the machine learning techniques. In the engineering view point, concentration is ambiguous and lacks clear definition. There is no established method to measure the degree of the concentration. We in this paper begins with a discussion of the concept of concentration, and then a discussion of how to measure it by using standard devices and sensors. In this study, we define the concentration degree over a short time interval of about five seconds. The value takes continues value from 0 to 1, and is determined based on the efficiency of simple work, which is typing tasks, performed in the interval. We convert the continuous value into discrete values: low, middle and high. The reason for discretization is for ease of use and understanding. Studies in [1] define concentration by using quantities measured by special devices such as the heart rate meters and the EMGs (electro-myography). These devices are not common in a daily life and are expensive so that we cannot use them in the usual classroom. For these reasons, we adopt the approach that uses only a simple Web camera having relatively low resolution.

1.1 Related Works

Many different methods [4, 5, 8] are proposed to estimate the emotions of a person. In particular, Google provides a tool that can estimates emotions from a facial image and other similar tools are available in the Internet. Emotion is usually treated as a combination of several basic elements, such as delight, anger, sorrow and pleasure. Typical tools estimate the proportion of these elements in a given face image. Note that there are different opinions how to determine these elements. On the other hand, although concentration and emotion include similar views, there is no established view of concentration. As far as the authors know, only a few general tools are usable for estimating concentration. Thus studies deal with concentration depending on the sensors and purposes. The methods in [11, 13] use special devices to get signal and analysis. In particular, accelerometers are often used to measure the moves of body parts, and then the concentration degree is estimated from movement patterns. Although this method is promising, there is still a problem of where and how many accelerometers must be attached to the body. Methods of using heart rate meters and electromyography, EMG, have also been studied [1]. The problem is that these sensors are not common for use in daily life. For these reasons, this study uses facial expression data from a Webcam that can widely use anywhere without cost issues.

2 Development of System and Experiment

2.1 Degree of Concentration

Concentration can be defined from many different perspectives. In a lexicographic sense, concentration is a process in which a person puts a lot of attention or energy into a particular task or activity. As we can see immediately from this definition, the essential property of concentration depends on the task or activity. In this paper, our definition is based on experience in everyday life, the efficiency of simple tasks are affected by concentration. We thus adopt a method of defining the degree of concentration based on the efficiency of simple tasks, and typing is used as a simple task.

The maximum task efficiency is calculated, as shown in the formula (1), by the number of characters that can be typed in a certain length of time interval. The current efficiency is defined in a similar way using the current number of the typed characters. The concentration degree is defined by the ratio of the maximum to the current as we show in the formula (2). It is important to note that the efficiency of typing also depends on worker's memory. The speed becomes faster when the characters to be typed are memorized. This problem is dealt with by devising the experimental method. Another essential issue is that typing is not an intellectual task, and does not require intellectual thought and decisions. A skilled worker in typing can do the task with a kind of physical reaction. We discuss this issue in the conclusion.

$$Typing\ speed = \frac{Number\ of\ letters}{Time} \qquad (1)$$

$$Concentration\ degree = \frac{Current\ typing\ speed}{Max\ of\ typing\ speed} \qquad (2)$$

2.2 Outline of System

Figure 1 shows an outline of the proposed system. The system has two main phases: model construction and estimation using the model. The dotted line shows a flow in the model construction. The model is constructed by using the method of supervised learning with the training data. In the estimation phase, procedures run along with the bold lines. We feed a sequence of face images over time intervals to the model, and then the system estimates the concentration degree over the time interval. In the both phases, a Web camera is the only sensor used in this system. The user does not need to wear any sensor with him/her.

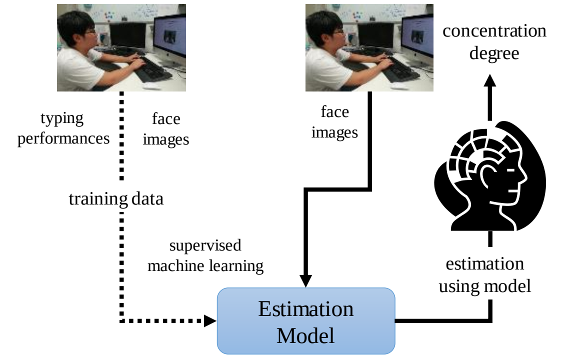

Figure 1: Outline of System

2.3 Collecting Training Data

Collection of training data is accomplished in the following three phases, and the collected data is used in the supervised machine learning process.

(1) Collecting typing data: In this phase, the efficiency of typing is measured to calculate the degree of concentration. This phase includes four different sections. In the first section, the participants use the input program to get acquainted with the keyboard as well as how to work with the system. In the second section, the participants enter the letters that appear on the screen randomly within 50 seconds. In order to eliminate the effects of memory, we give random characters in this section. In the third section, the participants enter the same way as they did in the previous step. However, in this section, the participants are asked to play word-chain games with another person. In the final section, the experiment participants continue to input with a condition. They have to answer addition, subtraction, and multiplication with random numbers no larger than 50. The last two sections are designed to intentionally change the concentration by external disturbances, which enables us to collect various degrees of concentration having different face expressions.

(2) Calculating concentration degree: The maximum performance value is stored for each participant, after trying several times of typing. Since we define concentration over a short time interval, we

divide the whole into 5-second intervals. Then using the equation (2), the values of concentration degree are in the range from 0 to 1. The value of 1 indicates the maximum efficiency, and 0 indicates no work. By observation, we divide data into three groups. The three groups includes almost the same number of participants as we show in Table 1. The first group is "high concentration" group, whose value is in [0.7, 1]. The second group is "middle concentration" having the value in [0.4, 0.69], and the other members, having the values less than 0.4, belong to "low concentration" group. The distributions of the three classes becomes almost equal, which is useful to eliminate the bad effect to the learning algorithm. These classes are associated with the corresponding images and used in the supervised learning.

(3) Collecting face images: We take video shots 5 frames per second simultaneously with the above typing data collection. Each concentration degree is calculated for each 5-second interval, all of the corresponding 25 images over the interval are labelled with the same concentration degree.

(4) Calculating head moves: In order to track head movements, we extract the facial landmarks [7] of participants. The landmarks distributed on a face. We use three characteristic points, the center of nose, and two points in canthus. We calculate three values that represent x-coordinate, y-coordinate, and z-coordinate in three-dimension space. In each frame, the values of the head position in x-coordinate and y-coordinate are taken from values of the point in center of the nose and those values are determined in pixels. The value in z-coordinate are specified in centimeters. First of all, we calculate the distance between two points in canthus to measure the distance between two eyes of each participant by taking a photo of them at the distance of 50 cm from the webcam. We estimate the distances of participants' heads to the webcam by calculating the length of the straight paragraph connecting two points in canthus.

2.4 Training and Building Model

By using the training data explained above, we apply supervised learnings with the two types of neural networks. Both of them are based on the CNN (convolutional neural network) [6].

In the first CNN, we use only the data of face images by cropping the facial area in every frames, unnecessary parts are simply removed. All of the images are resized into 500 x 500 pixels, and they are converted to grayscale ones before feeding them into the CNN. We here use the well-known VGG16 model [2, 10], which consists of 13 of 3 x 3 convolution layers and 3 pooling layers. Convolution layers change the resolution from 64 to 512 through 128 and 256. An input for this CNN is 25 images corresponding to the 5 frames of 5 seconds, whose labels are the discretized concentration degree. In the fully connected layer, we use the standard Softmax [6] function to get the output, which takes one of three labels: high, middle and low concentration.

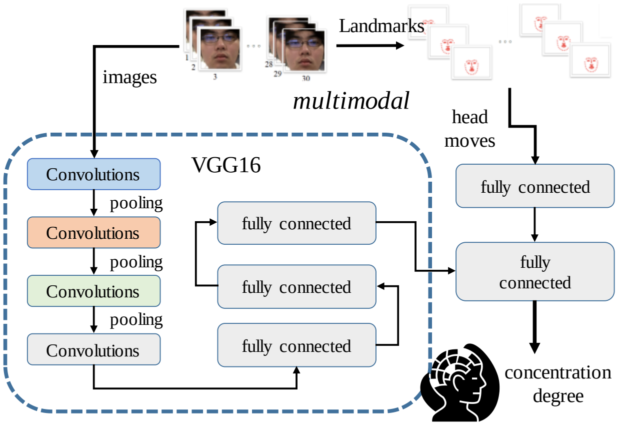

Figure 2: CNN with 2 Branches

The second CNN, shown in Figure 2, takes two types of input data; face images and head moves. The head moves is a time series data of real numbers as explained in the above. The first branch of this CNN is the same as the first one, so that it receives the only the same face images as the first CNN. The second branch receives the data of head moves. The values are sent to the fully connected layer in the figure, and then both branches. The two branches are joined into the final fully connected layer. Then the final output is obtained by using the standard Softmax.

2.5 Experiments

The purpose of the experiments here is twofold. First, we show the possibility of estimating the degree of concentration with only the data obtained through a simple camera. Second we show the possibility of using multimodal data, face images and moves in real numbers. The second type of CNN model is prepared to investigate this second purpose. There were total 10 participants in our experiments. All of them were male students, which ages from 21 to 27 years old, and they are not skilled worker in typing. The display and keyboard used here were not special but the standard ones. Each person was provided with a different account to log into the system. The experiments were carried out in a bright and comfortable environment, temperature and humidity were maintained within ranges that allow efficient work. There are methods to increase, by several operations on images, the training data, but they are not used here.

In order to confirm the validity of the degree of concentration by this method, all of the situation of the experiments was video-recorded, and confirmed by the subjects themselves. We use k-fold cross-validation to evaluate the prediction accuracy for unknown data. In this experiment, we set k=5 because the size of training data is relatively small. In the first experiment, the expected prediction accuracy for unknown data becomes 45%. The second experiment uses the facial images and the head moves. The prediction accuracy is improved to 62%.

3 Conclusions

This study shows that it is possible to estimate the degree of concentration, by using the machine learning techniques, with face images and derived data from them. The accuracy at present is relatively low. The main reason is that the number of training data is small and is not enough to reach a higher accuracy level. We do not use any techniques to increase data such as image deformation operations. We are planning to expand the training data and conduct experiments with it.

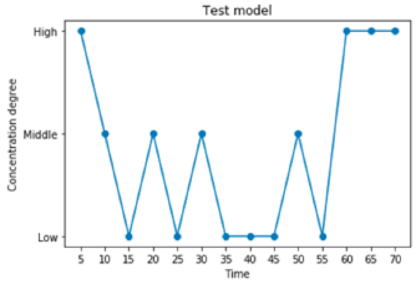

Figure 3: Changes over Time

The graph in Figure 3 shows the change of concentration for 70 seconds. Even if the accuracy in each 5-second section is not enough, it is considered to be effective to grasp the general situation for the whole. In particular, applying to the learning process of students, it is important to understand the situation in relatively longer time intervals. In this sense, short-term accuracy does not need to be extremely high. The degree of concentration defined this study is based on simple tasks without intellectual ability. It is a future task to verify whether this definition can be applicable to intelligent tasks. We also plan to calculate other motion patterns from the images, and to use them in the model. Blinks and fine vibration patterns of the bodies and faces are candidates of them.

We use the word-chain game and some easy calculation to force participants out of their focusing state. The shape of their mouths and heads might be different to their state in normal learning time. We are thinking about a new method that does not effect on their facial shape but could still make them out of focusing state, for example, allowing them to listen some kind of music while typing or changing the environment around them such as temperature, humidity, etc. In this way, we expect that we can get more natural data from participants to improve our system.

We are going to build our system in raspberry pi 3 to confirm that it is able to run in real-time with a weak computer. For the current scale of data, this environment provides enough performance for a practical use. We show the possibility of practical applicability in everyday learning environment.

References

[1] Asma Ben Khedher, Imène Jraidi, Claude Frasson, "Predicting Learners' Performance Using EEG and Eye Tracking Features", *The Thirty-Second International Florida-Artificial Intelligence Research Society Conference (FLAIRS-32)*, 2019.

[2] Ran Breuer, Ron Kimmel, "A Deep Learning Perspective on the Origin of Facial Expressions", [Online]. Available: https://arxiv.org/pdf/1705.01842.pdf.

[3] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, Pierre-Alain Muller, "Deep learning for time series classification: a review", [Online]. Available: https://arxiv.org/pdf/1809.04356.pdf.

[4] A. Godfrey, R. Conway, D. Meagher, G. ÓLaighin, "Direct measurement of human movement by accelerometry", Medical Engineering & Physics, Elsevier, Vol.30, No.10, 2008.

[5] Yelin Kim, Tolga Soyata and Reze Feyzi Behnagh, "Towards Emotionally Aware AI smart Classroom: Current Issues and Direction for Engineering and Education", *IEEE Access*, Volume 6, 2018.

[6] Shan Li and Weihong Deng, "Deep Facial Expression Recognition: A Survey", [Online] Available: https://arxiv.org/pdf/1804.08348.pdf.

[7] Daniel Merger, Matthias Rock, Gerhard Rigoll, "Robust Facial Landmark Detection via a Fully Convolutional Local-Global Context Network", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Masashi Okubo and Aya Fujimura, "Development of Estimation System for Concentrate Situation Using Acceleration Sensor", *International Conference on Human-Computer Interaction (HCI)* , 2009.

[9] Reinhard Pekrun, Stephanie Lichtenfeld , Herbert W. Marsh, Kou Murayama, Thomas Goetz. "Achievement Emotions and Academic Performance: Longitudinal Models of Reciprocal Effects", [Online]. Available:https://www.researchgate.net/publication/313447133.

[10] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *International Conference on Learning Representation (ICLR),* 2015.

[11] Ronny Stricker, Sebastian Hommel, Christian Martin and Horst-Michael Gross, "Realtime User Attention and Emotion Estimation on a Mobile Robot", *55th International Scientific Colloquium*, 2010.

[12] Carlos Valiente, Jodi Swanson, Nancy Eisenberg, "Linking Students' Emotions and Academic Achievement: When and Why Emotions Matter", [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3482624/.

[13] Janez Zaletelj, "Estimation of Students' Attention in the Classsroom From Kinect Features", *10th International Symposiun on Image and Signal Processing and Analysis*, 2017.