Feasibility study on evaluation of audience's concentration in the classroom with deep convolutional neural networks

Ryosuke Yoshihashi

Daiki Shimada

Hitoshi Iyatomi*

Hosei University
Dept. Applied Informatics
Tokyo, Japan
* iyatomi@hosei.ac.jp

***Abstract*—In this paper, we developed an estimation system for degree of audience's concentration by estimating individual's behavior with a deep learning approach. Our system firstly detects candidate location of audiences (CLAs) from the movie with Ada-boost classifier composed of Haar-like filters and their integration process. Then, each CLA is investigated to determine the target audience is "concentrated", "not concentrated" or "no exist" with 5-layered deep convolutional neural networks (DCNN). We used a total of 13 movies of which 3 movies were used for training of DCNN and the remains for evaluation. Our system achieved audience detection performance of precision = 84.8% and recall = 61.8% and estimation accuracy of individual attention as 72.8%.**

Keywords—faculty development; image analysis; convolutional neural network

I. INTRODUCTION

Quantitative evaluation of the degree of audiences' concentration is not only helpful feedback for speaker but might also one of the objective indicators for appealing of the lecture or presentation. Further, availability of objective estimation of individual's behavior would be more meaningful especially in the classroom; it makes lecturer to have an opportunity to follow ones who have low attention.

So far, faculty development program has been widely conducted in academic organizations to improve mainly their education effects [1-3]. One of the common methodologies of these activities is to have a questionnaire during the semester. However, it is fairly costly and the evaluation of lectures and their comparisons are difficult because each feed-back from a student is subjectively made in most cases.

In common, sensory devices such as electroencephalography [4,5], combination of several sensors [6] have been used for evaluation of subject's condition or behavior. These techniques are promising for limited situations, however they have a difficulty in applying in general situations due to their cost and limitations. For the abovementioned and its similar purposes, several methods based on image processing techniques have been proposed. Saida et al. [7] proposed a novel procedure to estimate the attention of students in the classroom with investigating the differences among image frames. This method did not identify individual and yields only the summarized "degree of attention" in the classroom. Lee's method [8] estimated the concentration of an individual student in the classroom. This achievement is promising, while size of the room is limited.

Chen [9] proposed a real-time estimation system of students' intentness. However, it requires a camera per person so that it is highly expensive.

In general pattern recognition problems, obtaining efficient parameters (or sometimes pattern patches) for the target task is considerably important. The "recognition" here means that it obtains the relationship between those extracted parameters, not the observed pattern itself, and the classification classes (e.g. labels). Since the collecting of the efficient parameters requires prior knowledge about the task and repetition of trial-and-error, it is unfortunately usually difficult. Our facing task is not easy, too. We need to tackle with crowded targets in the room simultaneously, each has a wide variety of appearances, posture and may under different lighting conditions, and not small numbers are overlapped each other. We need to collect quite robust image features over those various kinds of disturbances.

Several approaches addressing these issues have been proposed. In image recognition task, many methods attempt to extract robust features such as SIFT (scale invariant feature transform), HoG (histogram of oriented gradient), Gabor filter etc. and then perform the classification based on the model trained by supervised learning.

While on the other hand, so-called deep learning approach aimed to solve abovementioned issue has received increasing attention. LeCun et al. proposed epoch-making neural network model called "convolutional neural network" (CNN) [10] consisting of 8 layers networks; input layer, five convolutional and sub-sampling layers, one full connection layer and output layer. CNN attains the direct relationship between the image and the class owing to its capability of automated acquisition of local and sophisticated characteristics by means of multi-layered convolution layers and the following full-connection layer trained by gradient-based supervised training.

One of the most difficult issues in attempting multi-stratification of conventional multi-layered neural network such as back-propagation (BP) to adapt higher and large scale task is the over-fitting due to its large degree of freedom. LeCun's network successfully eliminated over-fitting by introducing sub-sampling layer and asymmetry property of the connection weights.

978-1-4799-7672-0/14/$31.00 ©2014 IEEE

Recently, Krizhevsky et al. [11] proposed the deep convolutional neural networks (DCNN) based on the CNN. It consists of a total 9 layers; input layer, five convolutional layers in some of which with local normalization and pooling, two full connection layers and output layer. DCNN is introduced several state-of-the-art ideas such as rectified linear units (ReLU) as an activating function of neuron, local response normalization [12], overlapping pooling, drop-out technique [13] which randomly enforces the weights zero to help elimination of over-fitting. DCNN achieved lowest top-5 error rate of 15.3% at the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC-2012), whose dataset is composed of 1000 class, each has around 1000 images.

DCNN had proven its quite high generalization and classification ability and therefore we decided to introduce this model for our task.

II. ESTIMATION SYSTEM FOR DEGREE OF CONSENTRATION

The proposed system receives a movie of the audiences as the input and outputs estimated degree of concentration of each individual. The schematic of the system is illustrated in Fig.1. The system is composed of (1) the audience detection part and (2) the attention estimation part.

The audience detection part operates for every pre-definitive period of the time and detects candidate location of audiences (CLAs). The following attention estimation part uses the trained DCNN and yields either of “concentrating”, “not concentrating” or “no audiences exist” for each CLA.

In this study, we cannot estimate the “actual” status of each audience, so we subjectively defined the status as follows: “concentrating” is the state that audience turned to speaker and have the posture trying to hear the talk. : “not concentrating” is the status not above.

A. Audience detection part

In this part, rectangular CLAs are determined by detecting upper body of audiences using Adaboost algorithm with Harr-like patterns [14]. Since inappropriate CLAs can be eliminated in the following part, this part makes much account of no omission in detection rather than their accuracy.

Appropriate determination of CLAs is not easy because audiences are often overlapped and each of which has wide variety of appearances, posture and may under different lighting conditions. In order to cope with this issue, our method detects rectangle pre-CLAs from image frame every τ second during the pre-defined period T seconds, and performs their integration process to form CLAs. In our experiments we experimentally decided τ = 30 and T = 300; total 10 image frames were used to determine CLAs. The detailed integration process is as follows:

Let arbitrary pre-CLA detected in two successive image frames with the interval τ namely i and j, their gravity center (x_i, y_i) and (x_j, y_j), and their area s_i and s_j. When both of the following two conditions are satisfied, we consider those two pre-CLAs focuses on the same audience and are integrated.

$$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} < \theta_d \tag{1}$$

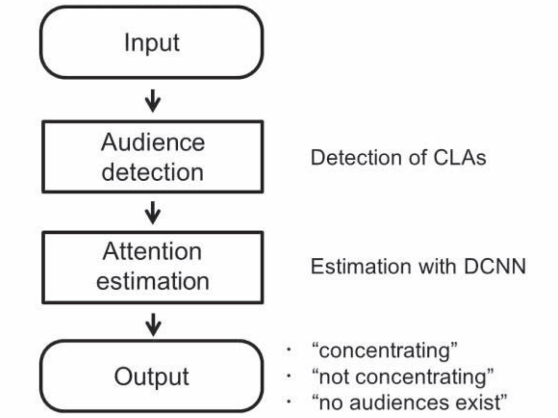

Fig.1. The schematic of the system

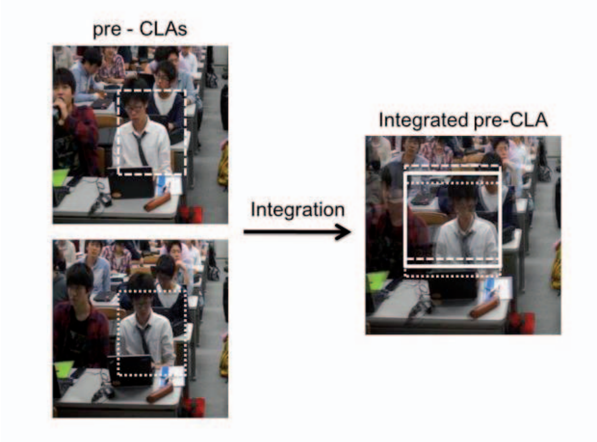

Fig.2. Integration of pre-candidate location of audiences (pre-CLAs)

$$\theta_S > \left|\frac{s_i}{s_j} - 1\right| \tag{2}$$

Here, θ_d and θ_S are thresholds and determined by pre-experiments. The integration process forms new pre-CLA, k, having the averaged gravity center, width and height of pre-CLAs of i and j, respectively. Fig. 2 shows an example of the integration process. The integration process performs these steps recursively to determine the location and size of CLAs.

B. Attention estimation part

Our method evaluates each CLA with trained deep convolutional neural network (DCNN). The schematic of our DCNN is shown in Fig.3. Our DCNN has 3 convolutional and input / output layers without any full connection layers. The number of layers and their size are determined by our pre-experiments. The summary of the network structures are in Table I. DCNN usually has a few full connection layers after the convolutional layers so as to tackle a big issue, while introducing them for our task deteriorates the performance due to over-learning caused by not so large data size. Therefore we did not use fully connection layers. Note that output of the last convolutional layers and the output layer are fully connected. The detailed explanation of each layer is in the followings.

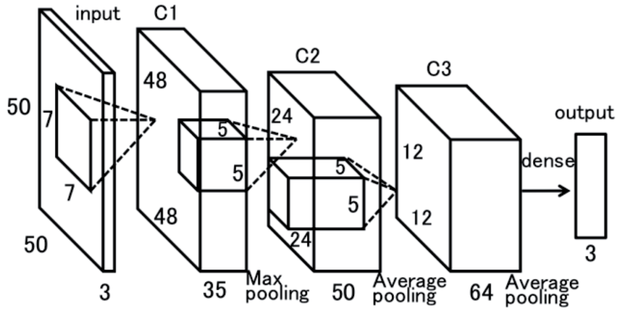

Fig.3. Schematics of our DCNN

TABLE I. THE ARCHITECTURE OF OUR DCNN

ID	conv size*	map size	#map	LCN†	pooling	other‡
input	7x7	50x50	3	-	-	
C1	5x5	48x48	35	3x3	max3x3	padding=2
C2	5x5	24x24	50	3x3	ave 3x3	padding=2
C3	-	12x12	64	-	ave 3x3	
output	-	3	1	-	-	

*: size of convolution window

†: local contrast normalization

‡: stride size of convolution window in convolutional layers is set to 1.

a) Input layer

The input layer consists of 3 channel 50x50 neurons (total of 7500 inputs); each corresponds to each pixel of CLA determined in the audience detection part.

b) Convolutional layers (C1-C3)

Our DCNN has three convolutional layers, namely C1, C2 and C3. Convolutional layer performs convolution to its former layer with multiple local "kernels" (i.e. convolutional functions) so as to map the local features of the former layer to the convolutional layer. Multiple kernels bring in the manifold of the model.

Now consider the convolution process between the input and the C1 layers. We prepared 35 7x7x3 kernels here. Each kernel in convolution layer convolves the local region in the former layer (i.e. input layer) and its successive planer scan makes one image plane called "map" or "feature map" (i.e. Accordingly 35 maps are formed in the C1 layer).

Assume an arbitrary neuron of the input layer $I_{x,y,c}$ which receives the input image pixel at the location (x, y) and color channel $c = \{R, G, B\}$.

The value of the neuron at the location (x, y) in the m-th map of C1 layer $\mathrm{u}_{x,y}^{C_1\_m}$ is calculated as follows:

$$\mathrm{u}_{x,y}^{C_1\_m} = f\left(\sum\nolimits_{\phi} w_{x,y,c}^{m} I_{x,y,c} + w_{0,0}^{m}\right) \tag{3}$$

Here, $w_{x,y,c}^{m}$ is the m-th convolutional kernel, $w_{0,0}^{m}$ is the m-th bias, ϕ is the convolution scope (i.e. 7x7 with (x, y) as its center) in the C1 layer, respectively. $f(.)$ is the non-linear activation function, called rectified linear unit (ReLU) as follows:

$$f(x) = max(0, x) \tag{4}$$

This function achieved better performance than conventional sigmoid type one in the original DCNN manuscript [11]. In our network design, we introduced padding on the perimeter of the map in the C1 and C2 layers so as to make their information located around the verge of map easier to transmit to successive layers.

c) Local normalization (C1, C2)

After the convolution process in the layers C1 and C2, local normalization [12] was conducted on each map. This technique is applied in several models and performance improvement with it has been reported.

d) Pooling (C1-C3)

We conducted "pooling" as the last process in each convolutional layer. Pooling is the process that calculates candidate value of local region in each map. While this process reduces the amount of overall information, it contributes to improve robustness over position gap. Max-pooling is conducted in the C1 layer and average pooling is done in the C2 and C3 layers with pooling size of 3x3 and pooling stride of 2, respectively.

e) Output layer

All neurons of the output layer and the C3 layer are fully connected as explained earlier. The final output of the n-th neuron in the output layer ($O_n; n = 1,2, \dots, N$) is calculated as follows.

$$O_n = \frac{\exp(y_n)}{\sum_j \exp(y_j)} \tag{5}$$

Here, y_j is the j-th input signal$(j = 1,2, \dots, N)$ to the output layer from the C3 layer.

f) The training of DCNN

The training targets of the DCNN are all kernel variables in convolutional layers, connection weights and biases. They are randomly initialized and iteratively updated with back-propagation method based on stochastic gradient descend.

III. RESULT

A. Material

In our experiment, we used a total of 13 movies of students recorded from different lectures. As for the training data, we captured 3 movies at 1 fps and determined CLAs. We assigned a training signal ("concentrating", "not concentrating" or "no audiences exist") to a total of 23,444 CLAs from arbitrary selected sequential 1891 image frames. Note that around 60,000 CLAs were determined from those image frames, however ambiguous cases were excluded in this experiment. In 10 evaluation movies, we randomly selects image frames with a time period T (300 sec) in which captured interval of τ (30 sec) and accordingly 10 image frames were selected from each movie. A total of 6,160 CLAs from 100 image frames were determined and inputted to the trained DCNN for evaluation. In these 100 image frames, the number of total

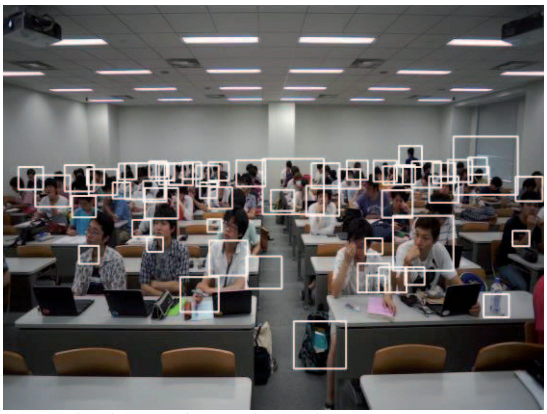

Fig.4. Example of detected CLAs in the classroom

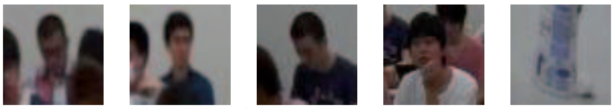

Fig.5. Example of detected each CLA

TABLE II. THE SUMMARY IN ESTIMATION OF THE ATTENTION OF AUDIENCES WITH DCNN

		Prediction		
		consent.	***no-consent.***	***no exist***
Actual	***consent.***	809	373	61
	no-consent.	559	1415	120
	no exist	116	449	2258

students should be identified was 5110.

B. Detection of students and estimation of the attention

The example of detected CLAs (white boundary boxes) in the audience detection part is shown in Fig. 4 as well as those of individual CLA is in Fig. 5. The confusion matrix of final results of our system is shown in Table II.

We found several errors in detecting students in this part especially in the front row and the backward position with highly overlapped, while we can confirm most of students are appropriately extracted. Note that several CLAs were excluded in the following part with the trained DCNN.

According to Table II, we confirmed our system detects students with the precision = 84.8% and recall = 61.8% ((809+373+559+1415)/5110; the total number of students). The accuracy of the three class classification was 72.8%, where the sensitivity of "concentrating", "not concentrating" and "no exist" were 65.1%, 67.6% and 80.0%, respectively.

IV. DISCUSSION

A. Estimation performance

In this study, we developed an estimation system of the degree of audiences' concentration only with their appearances (i.e. video images) with the hope that it fits instractor's visual perception.

In the audience detection, our system had a difficulty in detecting students who sit in front seat. This can be shown from low recall rate on detecting students (61.0%). While on the other hand, once students were determined, DCNN almost correctly discriminated the existence of the students, i.e. classification accuracy of existence of students based on detected CLAs, recall rate increased to 94.6%. This indicates DCNN accurately identify the existence of students with 50x50 image patch. From this comparison, we realized that improvement of the audience detection part was required.

In the attention estimation, the sensitivity (recall) rate for students' attention was limited; around 65-70%. In this study, we excluded CLAs in which multiple students photographed caused by their overlap in the training dataset, while they of course existed in the test set. Almost 30% of CLAs in the test set were applied in this case. Our DCNN had not trained these cases at all and therefore it yielded unsure output.

From this issue, we also reach the same issue that we need to improve the audience detection part for improving overall system performance.

B. Issues for further study

In this study, we estimated the degree of students' attention with only one captured image as a preliminary setting. We are sure we need to estimate them with monitoring a certain period of behavior of the students. We are also planning to develop a real-time estimation system in near future. It would provide lecturers to have an opportunity to improve their lectures according to the system output. We will investigate above issues to improve overall system performance to meet our system to be practical use.

V. CONCLUSION

In this study, we conducted feasibility experiment on evaluation of audience's concentration in the classroom with deep convolutional neural networks. Our system detects audiences with the precision = 93.4% and recall = 61.0% and the accuracy of the three class classification is 73.7%. We confirmed our system has a capability to evaluate audience's concentration effectually. We will improve the accuracy of audience detection and use plural frames for estimating audience's concentration in near future.

REFERENCES

[1] J. Froyd, J. Layne, D. Fowler, and N. Simpson, "Patterns for faculty development," Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual, pp.T1J-1 - T1J-5, 2007.

[2] J. Layne, J. Froyd, N. Simpson, R. Caso and P. Merton, "Understanding and improving faculty professional development in teaching," Frontiers in Education, 34th Annual, Vol.3, pp. S1C/15 - S1C/20, 2004.

[3] L. M. Aleamoni, "Faculty Development Research in Colleges, Universities, and Professional Schools: The Challenge," Journal of Personnel Evaluation in Education, Vol. 3, pp.193-195, 1990.

[4] K. Kawano, M. Yamamoto, M. Hirasawa, H. Kokubo, and N. Yasuda, "EEG analysis of children while concentrating on tasks," Journal of International Society of Life Information Science, Vol.15, No.1, pp.109-114, 1997.

[5] P. R. Davidson, R. D. Jones, and M. T. R. Peiris, "EEG-Based Lapse Detection With High Temporal Resolution," Biomedical Engineering," IEEE Transactions, Vol.54, No.5, pp.832-839, 2007.

[6] Y. Nishimura, Y. Tobe, "Suggestion of simple wireless sensors measure the concentration of teaching," IEICE 2011 (in Japanese).

[7] T. Saida and H. Yanai, "Discrimination of the levels of concentration of class participants with image processing," IEICE technical report.

Human communication science, Vol.103, No.742, pp.83-87, 2004(in Japanese).

[8] H-C. Lee, C-L. Wu, and L-J. Chen, “A Crowdsourcing-Based Approach to Assess Concentration Levels of Students in Class Videos,” Technologies and Applications of Artificial Intelligence (TAAI), pp.228-233, 2013.

[9] M. Chen, “Visualizing the Pulse of a classroom,” Proceedings of the eleventh ACM international conference on Multimedia, pp.555-561, 2003

[10] Y. LeCun, L. Bottou, Y. Bengio and P.Haffner, “Gradient-basedlearning applied to document recognition,” Proc. IEEE, Vol.86, No.11, pp.2278-2324, 1998.

[11] A. Krizhevsky, I. Sutskever and G. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems 25 (NIPS2012), pp.1106-1114, 2012.

[12] K.Jarrett, K.Kavukcuoglu, M. Ranzato and Y.LeCun, “What is the Best Multi-Stage Architecture for Object Recognition?,” Proc. 12th IEEE Conf. Computer Vision (ICCV) 2009, pp.2146-2153, 2009.

[13] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” The Computing Research Repository (CoRR), Vol. abs/1207.0580, 2012.

[14] H. Kruppa, M. Castrillon-Santana and B. Schiele, "Fast and Robust Face Finding via Local Context,” Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.157-164, 2003.