Automatic Engagement Level Estimation of Kids in a Learning Environment

Woo-Han Yun, Dongjin Lee, Chankyu Park, and Jaehong Kim

***Abstract*—This study is about the automatic engagement level measuring system which extracts features of kids taking tests with a desktop computer and estimates an engagement level. We recorded 12 kids from two difference kindergartens for 5 days. The test consists of 6 subjects with 2 sessions (levels). The recorded RGB video data is divided into 30 second video clips which are labeled by an expert. Cues reflecting face and head information are extracted from video data. The cues are aggregated for 30 seconds and used for estimating the engagement level. We used a relevance vector classifier to estimate an engagement level. We also analyze the data using linear regression analysis and find valid features. The system shows a promising performance of engagement level estimation of kids.**

***Index Terms*—Engagement level estimation, facial expression, facial feature, kid, multiple intelligence.**

I. Introduction

A task to measure an engagement level of people is an interesting research area in an affective computing field. The engagement level estimation technology has a range of applications. In the educational area, it is important to identify the engagement level of students. This engagement level could be used to give an appropriate feedback to the students such as a refresh time for bored students or more challenging task for interested students. For an advertisement or a market research purposes, the attention level of viewers is a main key to measure the effectiveness of advertisements. This attention level of viewers is valuable materials to make the next contents for the producers or gauging the effectiveness of advertising.

In this study, we focus on the engagement level measurement of kids in an educational environment, especially who take tests with interactive contents using a desktop computer. During the test, a video camera on the top of the monitor records a kid's face and head. We recorded 12 kids from two different kindergartens for 5 days. The subjects were selected among multiple intelligence theory [1]. Every subject consists of 2 sessions (levels). The total collected database consists of 144 videos (6 subject × 2 sessions × 6 kids × 2 kindergartens). The collected video data is divided into 30 seconds video clips. From these video clips, we extract engagement cues in a frame-by-frame basis and aggregate these cues into features. We construct the system with these features to predict the engagement level of kids.

The rest of this paper is organized as follows. First, we review the related works in an engagement level estimation. Next, we describe the database construction process including environments and annotations. We, then, explain the candidate cues extracted from face and head in a RGB video. We describe the aggregation method and decision process. In the experiment section, we show the experiment results and evaluation analysis. Finally, we conclude our work.

Manuscript received August 7, 2014; revised December 9, 2014. This work was supported by the R&D program of the Korea Ministry of Knowledge and Economy (MKE) and the Korea Evaluation Institute of Industrial Technology (KEIT) [10041826, Development of emotional features sensing, diagnostics and distribution s/w platform for measurement of multiple intelligence from young children].

Woo-han Yun, Dongjin Lee, Chankyu Park, and Jaehong Kim are with Human-Robot Interaction Section, Electronics and Telecommunications Research Institute, Daejeon, South Korea (e-mail: yochin@etri.re.kr, robin2002@etri.re.kr, parkck@etri.re.kr, jhkim504@etri.re.kr).

II. Related Works

Studies related to estimating the engagement levels or the affective states of people have been conducted in several research groups. There was a feasibility study to measure the engagement level of TV viewers using face and head gestures [2]. The authors showed the automatic engagement recognition is possible in a naturalistic environment with low computation cost algorithm and non-invasive sensors such as RGB video cameras. They used head orientation, face distances/angles, head roll, head size and position. McDuff et al. presented a system, AffectAura, which predicts a computer user's affective state such as valence, arousal, and engagement using audio, visual and physiological data and a user's log [3]. They showed this system is useful for users to reconstruct their stories in their own memories. There was a trial to build an affective recognition system for kids in a learning environment [4]. The authors utilized face and head gestures from video data, posture features from a pressure sensing chair, and features from task information. They used a Gaussian Process to combine this multiple features.

There are various works for estimating the engagement level or affective state from visual cues. Our work has differences in a several aspects. First, we only utilize the visual data from a RGB web camera. Even though there are a range of sensors, a RGB camera still popular sensing device because of its cheap price and non-invasive feature. Second, our work is focusing on kids. Even though kids are in a restricted learning environment, they act very naturally. They often move fast and go out of view of cameras. They were restless all the time during the test. In our work, we construct a robust module to cover these kids' features. Finally, we utilized a relevance vector classifier (RVC) for engagement level estimation [5]. RVC has several advantages over the classifiers used in the previous studies [2], [4] and yields more accurate results.

III. Data Collection

A. Environmental Setting

We first construct the experimental environment. Kids are sitting on the chair and taking a test. They solve several questions using a mouse while watching a monitor on the desk. The RGB video camera mounted on the top of the monitor records the face and upper body of kids with a resolution of 640 × 480 pixels at 30 frames per seconds. A Kinect camera, Galvanic Skin Response Sensor, and Photo Plethysmo Graphy Sensor are also equipped for other tests, but those are not used in this work. Fig. 1 shows the experimental environment.

Fig. 1. Experimental environment for data collection.

The subjects of the test were chosen from the abilities in the theory of multiple intelligences proposed by Howard Gardner [1]. We utilized a commercial testing interactive software [6]. The subjects of the test are to gauge the abilities of six categories: musical-rhythmic and harmonic ability, visual-spatial ability, verbal-linguistic ability, logical-mathematical ability, interpersonal ability, naturalistic ability, and existential ability. Two subjects, intrapersonal ability and bodily-kinesthetic ability, are not selected because those have a trouble to gauge the ability in a recorded video in a natural way.

The RGB video data was recorded in two different kindergartens for five days on each kindergarten. On each kindergarten, six kids participated in this test. Every subject is made up of two sessions, low and high level. The time of each session varies from 6 minutes to 17 minutes that depends on the subject and the level.

The contents of tests were organized for kids to take a test by themselves. Most of the kids take a test without an intervention of the teacher during the test. However, some kids who could not understand the question were guided by the teacher.

B. Annotation

The engagement level of each kid in videos was annotated by a human coder majoring in pedology. The video was divided every 30 seconds clips and labeled into 4 levels, high/low interest and low/high boredom. After the annotation, the criterion of labeling four engagement levels was reported and it is in the below.

1) High interest

- *Facial expression*: bright and earnest face, shining eyes
- *Eyes*: fixed eyes *on* the screen and concentrative
- *Posture*: correct *posture* nearby the monitor
- *Action*: no unnecessary actions

2) Low interest

- *Facial expression*: expressionless face
- *Eyes*: mostly fixed eyes on the screen
- *Posture*: mostly correct posture
- *Action*: no unnecessary actions

3) Low boredom

- *Facial expression*: tired face and grimace
- *Eyes*: looking off from the screen one or two times
- *Posture*: leaning backward or bending body
- *Action*: unnecessary actions including touching a part of the body

4) High boredom

- *Facial expression*: tired face and grimace
- *Eyes*: often looking off from the screen
- *Posture*: losing their posture or leaving the place
- *Action*: unnecessary actions including touching a part of the body or yawning

The total number of clips is 2,745. Each engagement level has each characteristic. These annotated labels are used as a ground truth for training and testing of the proposed method. The mainly considered behavior is visual focus of attention, facial expression, and body pose, but limited to these behaviors.

IV. Engagement Level Recognition

A. Cues

We extract the cues to reflect the face and head gesture information. This is because the face and head are mostly visible regardless of a distance and could be used in more general cases [2]. Face detection module plays a major role in face and head gesture information extraction. In case of kids, they act naturally such as tilting their faces, looking at other places, and twisting their bodies. In order to extract cues robustly in these cases, we combined the results of two differently trained face detectors and integrated these with face tracker to prepare against the failure of face detection. The flowchart of our face detection and tracking modules is in Fig. 2.

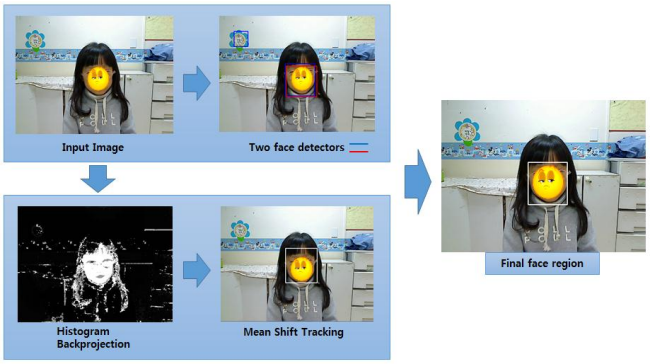


Fig. 2. Process to detect the face region of the kid using two face detectors and a face tracker.

From the face detection module, we extract following cues.

- **Face size and position:** We are *able* to extract the face size and position from the result of face detection module. We expect the size and position of a face could *capture* a relative distance and a location of kids from the monitor screen.
- **Head pose**: Head pose is another cue to determine the focus of a person's attention [2]. Because of our work assumption *that* the kids are sitting and solving the problem during the test, we could assume that undetected face could be considered as looking at other places. Head pose consists of two states: nonfrontal face and frontal face. If a frontal face is detected, it means the kids are looking at the monitor. Otherwise, we consider that the kids are looking off from the monitor.
- **Head roll**: We also calculate the absolute angle of the head using the position of eyes and horizontal plane with face detection and eye detection module. This head roll cue could capture the tilting head of the kids caused by someone's confusion or boredom.
- **Facial expression**: Facial expression is another widely used cue [2]-[4]. Among the facial expressions, the mostly used expression is smile. In our preliminary test, most of the kids did not show the expressions except for 'neutral' and 'smile'. They expressed 'smile' even though the situation where they are 'frustrated' or 'angry'. Facial expressions may not reflect the person's emotional state because of other factors such as a relationship with others and an atmosphere [7]. Therefore, we used 'smile' or 'the others' as a cue from facial expression module [8].

After all above cues are extracted from the video data, we apply the median filter with window size 3 to the cues to remove the noise such as impulse noise (cause by an incorrect detection module) because the noise could ruin the result of aggregation such as max, min, range, and variance.

We also use the displacement of cues between adjacent frames as a cue by calculating the absolute difference between neighbor cues in a time space. This displacement could reflect the variations of cues along to a time.

The process to extract cues is in Fig. 3.

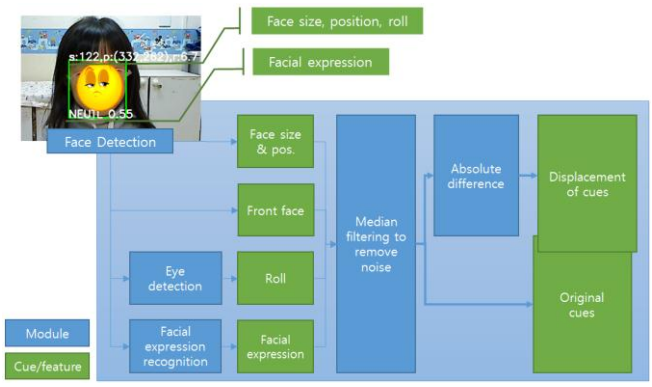


Fig. 3. Process to extract cues from a video data. A blue box and a green box represent a processing module and a result cues, respectively.

B. Aggregation

All cues including original cues and their displacements are extracted from every six frame (5 frames per second). These cues are aggregated using different functions over a 30 second time window (150 frames = 30 seconds × 5 frames per second). The functions we used find the following values: minimum value, maximum value, mean, median, standard deviation (STD), range (difference between maximum value and minimum value), and rate of zero crossings (ZCR, number of zero crossings/number of frames). We expect various features are extracted from video data with these cues and various aggregation functions. For example, the range of the face size may indicate the relative range of moving distance of a kid in the front and rear direction. The rate of zero crossings of a face size and a position may reflect how often kids move during the test. The cues and aggregation methods in this study were summarized in Table I.

TABLE I: MODULES, CUES (THEIR RANGES AND UNITS) AND AGGREGATION METHODS. IN A SECOND COLUMN (CUE), CUES IN AN ITALIC FONT MEANS DISCRETE SIGNALS AND CUES IN A NON-ITALIC FONT MEANS CONTINUOUS SIGNALS

Module	Cue (range and unit)	Aggregation Method	
Face detection	Face size (25 ~ 480 pixels)	**For continuous signals:** Mean, Median, Max, Min, STD, Range, ZCR	**For discrete signals:** Mean, STD, ZCR, Histogram
	Face position (13 ~ 468 pixels/y direction, 13 ~ 628 pixels/x direction)		
	Head pose (Front or not)		
Face & eye detection	Head roll (0 ~ 25.3 degrees)		
Facial expression recognition	*Facial expression (Smile or not)*		

* STD: Standard deviation
* ZCR: Rate of zero crossings

C. Dependency Removal

After aggregating the cues, the number of total features is 55 (34 from original cues and 21 from displacement of cues). In our case, some of the features have redundancy. For instance, the range of head roll is same with the maximum value of head roll because the min value of head roll is zero. This redundancy hinders a regression analysis and finding valuable features. We remove features having high redundancy by computing the correlation of each feature (correlation > 0.99). The number of remaining features is 47.

D. Classification

In this work, we used a relevance vector classifier. A relevance vector classifier (RVC) has several advantages. First, RVC is a sparse version of the Bayesian kernel logistic regression (which is also known as a Gaussian process classification (GPC)) [5]. RVC depends only sparsely on a training data by imposing a penalty for every non-zero weighted training example. This modification brings computational saving by using only selected data in an inference time and over-fitting prevention of the training set which has robustness to noisy data and generalizes better to new data. Second, RVC takes over the advantages of GPC such as preventing overconfident, assigning certainty to its class predictions, and modeling nonlinear relationship. RVC is similar with SVM, but RVC could result in more sparse solutions than SVM, accommodate more kernel functions, and be easily combined with other probabilistic models [5].

We compared RVC with other classifiers, a logistic regression, linear support vector machine (SVM), kernel SVM with a radial basis function (RBF), and GPC. We utilized LIBSVM library [9] for two SVMs and codes in a matlab for a logistic regression model.

V. Experimental Setting

A. Analysis of Features

We analyze the features and identify valid features (p-value < 0.05) among the remaining features using linear regression analysis. This task is just to identify the valid features statistically, not used in the further process. Table II shows the selected valid features and Fig. 4 displays the mean and variance values of valid features among four classes.

TABLE II: Valid Features (CUE + Aggregation), Their p-Values and Meanings Extracted through Linear Regression Analysis

Cue	Aggregation	p-value	Meaning
Face position Y	mean	0.0000	Average height of face relative to the screen
Head pose	mean	0.0000	Average time of watching a screen
Head pose	ZCR	0.0000	Frequency of looking around
Expression	mean	0.0000	Average time of expressing a smile
Expression	ZCR	0.0000	Frequency of changing expressions
Displacement of face size	Mean/median	0.0000/ 0.0011	Body movement in a forward and backward direction
Displacement of face size	STD	0.0063	Variation of body movement in a forward and backward direction
Displacement of face position X	mean	0.0000	Body movement in a left and right direction
Displacement of face position Y	Mean/median	0.0000/ 0.0186	Body movement in a vertical direction
Displacement of face position Y	Max	0.0001	Maximum value of body movement in a vertical direction
Displacement of face position Y	STD	0.0000	Variation of body movement in a vertical direction

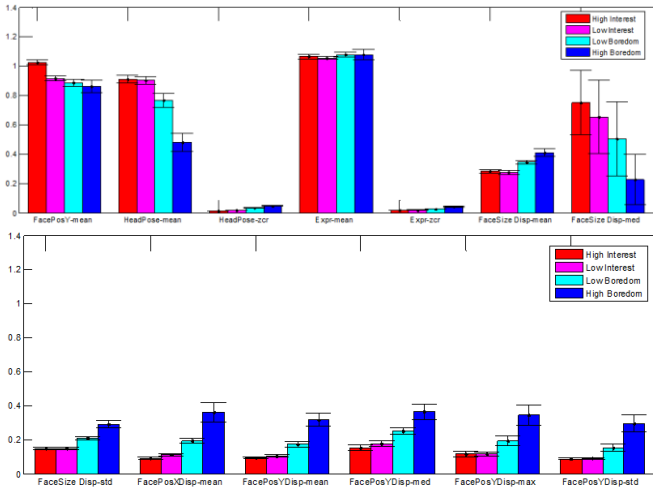


Fig. 4. Mean and variance of valid features of four classes.

In the Fig. 4, we could find that the displacements of cues are a good indicator to estimate an engagement level which coincides with our intuition. For instance, the mean of displacement of face position in *x* and *y* direction means how often kids move during the test. High mean value of displacement of face position in *x* and *y* direction may indicate that kids move frequently their bodies because they feel bored and could not concentrate on the test. The mean and ZCR of head pose means how long kids concentrate on the test and watch a screen. Low mean and high ZCR values of head pose may represent kids could not constantly concentrate on the task and look around.

B. Preprocessing

In the test, all cues are extracted in every six frames (5fps). We divided each video into smaller non-overlapping video clips (30 seconds). We applied aggregation methods for 30 second video clip (150 frames = 5 frames per second × 30 seconds). The total number of video clips in our work was 2,745 and average number of clips per kid was 228.7 video clips. The distribution of the class labels of our collected data is in Fig. 5. The labels of data are severely biased toward classes 'High Interest' and 'Low Interest'. We think this is because the kids have given a goal to solve a question and the test is also enjoyable interactive contents which entices the kids' attention.

We used two-fold cross-validation. This means that half kids of the data are used for test and the others for training. This test is repeated vice versa.

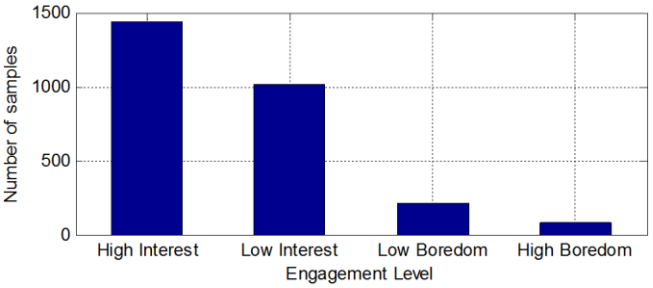


Fig. 5. Distribution of class labels.

In this study, we focus on two class problem, Interest and Boredom by considering High and Low classes into one. In training and test, we used different costs for each class because the number of data for each class is severely biased to the class 'Interest'. We used data from eight kids among twelve kids because four kids showed only (High and Low) Interest feedback.

C. Result

We tested six classifiers, a baseline classifier, RVC, GPC, a logistic regression, a linear SVM, and a kernel SVM using a RBF kernel. The baseline classifier is designed to show the improperness of the ordinary accuracy. The baseline classifier assigns one class which is the most common label in training set without considering input test data. The most common label is class 'Interest' in our case. The accuracy of these classifiers is in Table III. In the result table, we show two kinds of the accuracies, which are ordinary and balanced accuracy. A balanced accuracy is calculated to balance a different number of samples among classes by giving different weights on the test samples.

In the result, a baseline classifier assigning only class 'Interest' showed the best accuracy 84.28% which means the portion of class 'Interest' in a test set is 84.28%. In a balanced accuracy, the baseline classifier showed 50% which is a reasonable result. However in a real situation, we also consider the prior knowledge which class has much more chances to be found among all classes. Therefore we consider both types of accuracy. In ordinary accuracy, the baseline classifier, a logistic regression, and RVC showed higher accuracy. A kernel SVM, GPC, and RVC showed good performance in a balanced accuracy. RVC showed higher accuracy in both types of accuracy. Our result is slightly lower than the previous study [4] even though the experimental settings and scenarios are different. However, this result is still comparative because the previous study used data from several sensing devices such as a camera, a pressure sensing

chair, and other information from contents, but we only used a data from a RGB web camera.

TABLE III: ACCURACY OF THREE CLASSIFIERS

Classification method	Accuracy (%)	Balanced Accuracy (%)
Baseline classifier	84.28	50.00
Logistic regression	79.20	63.16
Linear SVM	73.85	65.21
Kernel SVM(RBF)	74.27	69.65
GPC	76.75	68.14
RVC	78.53	70.64

Fig. 6 shows the prediction of the balanced accuracy based on each feature using logistic regression. The features recording higher accuracy roughly coincide with the valid features in linear regression analysis.

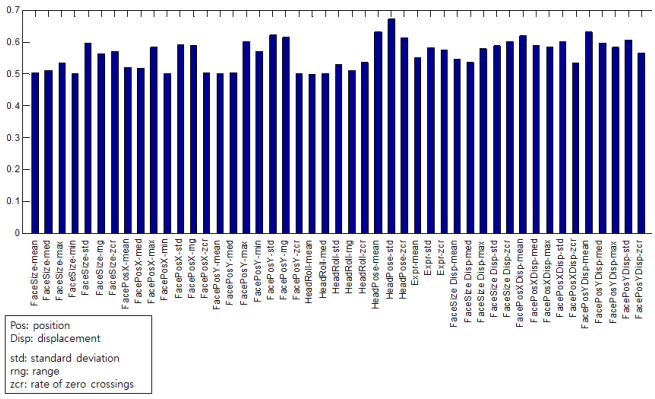


Fig. 6. Balanced accuracy of logistic regression using each feature.

VI. CONCLUSION

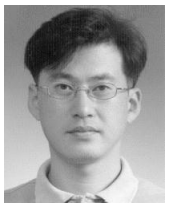
In this study, we showed an automatic engagement level measurement system for kids in a learning environment. We collected a database from 12 kids using a RGB web camera during the multiple intelligence test. The video was divided into 30 seconds smaller clips and extracted cues for every six frames. The cues are integrated for a small clip into one feature. We used RVC to classify features which showed the promising result. The accuracy was comparative when it is compared with other previous work even though the experimental setup is quite different. We also analyze the features using linear regression analysis and winnowed valid features.

REFERENCES

[1] H. Gardner, *Frames of Mind: The Theory of Multiple Intelligences*, Basic Books, 1983.
[2] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of TV viewers," in *Proc. the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1-7.
[3] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "AffectAura: An intelligent system for emotional memory," in *Proc. the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 849-858.
[4] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proc. the 13th Annual ACM International Conference on Multimedia*, 2005, pp. 677-682.
[5] S. J. D. Prince, *Computer Vision: Models Learning and Inference*, Cambridge University Press, 2012.
[6] Multiple Intelligence Institute Co. Ltd. [Online]. Available: http://www.multiiq.com
[7] M. E. Hoque, D. J. McDuff, and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 323-334, 2012.
[8] W. Yun, D. Kim, C. Park, and J. Kim, "Hybrid facial representation for emotion recognition," *ETRI Journal*, vol. 35, no. 6, pp. 1021-1028, 2013.
[9] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1-27:27, 2011.

Woo-Han Yun received the B.S. degree in electronic and electrical engineering from Sung Kyun Kwan University, Suwon, Korea, in 2004, and the M.S. degree in computer science and engineering from Pohang University of Science and Technology (POSTECH), Pohang, Korea, in 2006. He is a research scientist at ETRI (Electronics and Telecommunications Research Institute) since 2006. His current research interests include object recognition, face recognition, facial expression recognition, and affective computing.

Dongjin Lee received his M.S. degree in computer software engineering from University of Science and Technology (UST), Rep. of Korea, in 2013. He has been a research scientist at ETRI since 2013. He is also working toward his PhD degree in electronics at Chungnam National University. His research interests are in computer vision, machine learning, and affective computing.



Chankyu Park received the MS degrees both in electronics engineering from Kyungpook National University in 1997, and in software engineering from Carnegie-Mellon University in 2005. He joined Electronics and Telecommunications Research Institute (ETRI) in 1997 and has been a principal member of Engineering Staff in the Intelligent Robot Research Division since 2004. He is also working toward the PhD degree in electronics at KAIST. His research interests are in computer vision, machine learning, and bio-signal processing.

Jaehong Kim received his PhD from Kyungpook National University, Daegu, Rep. of Korea, in 2006. He has been a research scientist at ETRI, Daejeon, Rep. of Korea, since 2001. His research interests include socially assistive robotics for elder care, human-robot interaction, and gesture/activity recognition.