IEICE TRANS. INF. & SYST., VOL.E100–D, NO.2 FEBRUARY 2017

PAPER

# Learning State Recognition in Self-Paced E-Learning

**Siyang YU**[†a)], ***Nonmember***, **Kazuaki KONDO**[††], **Yuichi NAKAMURA**[††], ***Members***, **Takayuki NAKAJIMA**[†††], ***and*** **Masatake DANTSUJI**[††], ***Nonmembers***

**SUMMARY** Self-paced e-learning provides much more freedom in time and locale than traditional education as well as diversity of learning contents and learning media and tools. However, its limitations must not be ignored. Lack of information on learners' states is a serious issue that can lead to severe problems, such as low learning efficiency, motivation loss, and even dropping out of e-learning. We have designed a novel e-learning support system that can visually observe learners' non-verbal behaviors and estimate their learning states and that can be easily integrated into practical e-learning environments. Three pairs of internal states closely related to learning performance, concentration-distraction, difficulty-ease, and interest-boredom, were selected as targets of recognition. In addition, we investigated the practical problem of estimating the learning states of a new learner whose characteristics are not known in advance. Experimental results show the potential of our system.

***key words:*** *e-learning support system, learning states recognition, interpersonal differences, classifier selection*

## 1. Introduction

Self-paced e-learning is a widespread learning method that benefits from the characteristic of flexibility. Its learning performance is highly dependent on a learner's autonomy. Hence, well-designed self-paced e-learning should attract and motivate learners effectively. A learner's cognitive-affective states during self-paced e-learning provide significant feedback that can serve in many phases, including design, development, utilization, management, and evaluation of processes and resources for learning. However, obtaining such valuable information is difficult. Freedom from time and locale brings both convenience and difficulties. Therefore, acquiring such information is demanding.

An automated system that can assist teachers in recognizing students' learning states is required to deal with this problem. Such a system will considerably reduce the effort required by the teacher, especially when the number of students is large.

In this research, we designed an e-learning support system that can capture learners' behaviors visually and estimate their learning states in an actual self-paced e-learning environment. Then, we considered the practical problem of estimating the learning states of a new learner whose characteristics are not well known in advance. E-learning systems may face a variety of new students, because e-learning is often designed as learners with a variety of backgrounds can join at different time and places. However, it would be difficult to have all types of learner models beforehand. We need to consider a mechanism to adapt the system to new learners.

In the following sections, we first introduce our framework, our method for learning state estimation, and the scheme for dealing with new learners. Then, we present our experimental results showing good potential for our framework.

## 2. Related Work

Previous studies have focused on recognizing learners' cognitive-affective states in learning environments. These had different target states and utilized various modalities. Butko et al. [1] proposed an automatic facial feature extraction system that was designed based on the Facial Action Coding System. Seven GentleBoost classifiers were used to recognize the expression of being interested, thinking, tired or bored, confused, confident or proud, frustrated, and distracted on the part of a learner during interactions with a teacher. Ammar et al. [2] focused on detecting the contour of eyes, eyebrows, and mouth. Distance changes among these were used for classifying six universal emotions. Whitehill et al. [3] investigated the correlation between facial expressions and self-reported difficulty. Zakharov et al. [4] used facial features to identify whether the affective state was positive or negative. This enabled a pedagogical agent persona to respond to a learner's action on the basis of the learner's cognitive and affective states. D'mello et al. [5] studied dialogue features extracted from conversations between learners and an intelligent tutoring system and on the classification of boredom, confusion, flow, frustration, and neutrality. Litman et al. [6] used acoustic and prosodic features of students' speech to detect negative, neutral, and positive emotions. In [7], physiological signals were used to recognize emotions developed during the learning process. Three kinds of sensor were used for skin conductance, blood volume pressure, and electroencephalography (EEG) to recognize four kinds of emotion, engagement, confusion,

Manuscript received April 4, 2016.
Manuscript revised September 15, 2016.
Manuscript publicized November 21, 2016.

[†]The author is with Graduate School of Engineering, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The authors are with Academic Center for Computing and Media Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.
[†††]The author is with Graduate School of Human and Environmental Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.
a) E-mail: yusiyang@ccm.media.kyoto-u.ac.jp

DOI: 10.1587/transinf.2016EDP7144

boredom, and hopelessness. Yang [8] used combinations of mouse operations and facial information to detect attending and responding states of students, including attentive vs. inattentive and active vs. passive, respectively. Woolf et al. [9] combined four sensors to recognize confident, frustrated, excited, and interested, using facial expressions, postures measured by the pressure from the seat cushion and back pad, a learner's hand pressure measured by a special mouse, and skin conductance.

## 3. Problems and Objectives

Despite much research progress having been made with respect to cognitive-affective state recognition in learning environments, problems remain for practical applications.

Various cognitive-affective states have been recognized, but some of them were either not closely relevant to learning performance, or could not be used readily by teachers. In other words, research from the perspective of assisting teachers has received less attention. One noteworthy issue in practical application that must be considered is the applicability of modality and equipment for raw data measurement. The choice must provide rich information, but more importantly, must be easily integrated into practical e-learning environments without imposing additional constraints on learners or learning environments. Another critical problem for practical use, which has been less explored, is how to deal with new students. We need to consider that learners are considerably different in their behaviors, i.e., inter-personal differences among learners. Even if we achieve sufficiently good performance for a specific learner, we might not obtain good performance for another learner.

To handle these problems, we choose learning state targets that have been proven closely related to learning performance. The details will be provided in the next section. Teachers can use them to improve learning experience effortlessly. We choose visual sensing of learners, because it is non-intrusive and non-contact, and a small camera can be integrated easily with existing e-learning systems with current technologies. Based on this, we develop a reliable method to estimate learners' states. Moreover, the e-learning support system requires models that can be applied to a variety of learners. We investigate the inter-personal differences and explore methods for choosing an appropriate model for a new learner.

An overview of our scheme is illustrated in Fig. 1.

There are three principal components. At each e-learning site, an RGB-D (color and depth) camera is mounted on an existing e-learning system, as shown on the top left in the figure. A video of a learner's face and upper-body is captured by a Kinect camera in RGB-D format. Learning states are estimated through the module as in the bottom portion of Fig. 1. In this component, visual features are obtained through the processing of RGB-D images. Subsequently, they are input to Support Vector Machine (SVM) to recognize learners' internal states. Details will be given

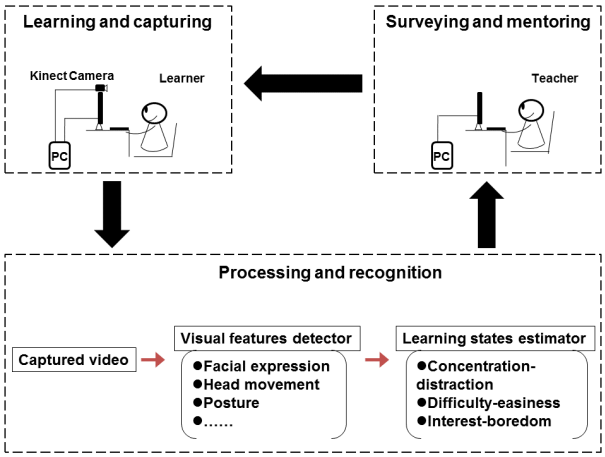


**Fig. 1** Overview of our scheme

in Sect. 5. The detected information is summarized and presented to teachers in a comprehensible manner, as depicted in the top right portion of Fig. 1. This information enables teachers to provide mentoring, modify materials, or conduct educational analysis. For instance, teachers can find and pay additional attention to those learners who need more assistance. Teachers can also design and adjust materials to keep learners motivated, e.g., by setting tasks a step ahead of learners' current skill levels. This use is left for future work, and we concentrate on the top left and the bottom portions in this paper.

## 4. Learning States

### 4.1 Internal States

We choose concentration-distraction, difficulty-ease, and interest-boredom as the targets of recognition. The significance of concentration in education has been investigated and emphasized in many studies, such as [10] and [11]. Difficulty of learning content plays a crucial role in maintaining students' concentration [12]. It is important for knowledge and skills acquisition to keep the tasks in the zone of proximal development of students by setting appropriate challenges. Interestingness is another important factor in motivating students to remain in concentration [13]. Moreover, a positive relationship between interest and academic achievement has been found in [14]. Shirey [15] reported that information can be learned readily, if it was interesting to learners.

It is widely recognized that they are mutually related. Concentration on a target can foster interest. Interest can be referred to as an important driving force which can result in concentration. However, concentration can be affected not only by interest but also by a variety of factors both inside and outside of learners, e.g., fatigue, stimulus strength, and time and place of learning. For difficulty and interest, Silvia [16] reported that both low and high in difficulty may cause low rating in interest, which implies that

**Table 1** Five-level scale of learning states

| 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|
| Very concentrated | Concentrated | Neutral | Distracted | Very distracted |
| Very difficult | Difficult | Neutral | Easy | Very easy |
| Very interesting | Interesting | Neutral | Boring | Very boring |

interest captures an aspect different from difficulty. Therefore, our scheme deals with those three as separate indices. Its advantage is clear if we think of their usage. Difficulty is an important feedback that is helpful for keeping the level of the learning materials adequate, e.g., learning materials need to be easier if learners feel too much difficulty and vice versa. Interestingness is also good information to make learning materials attractive. Teachers get good feedback how much learners are interested in each portion. Concentration is helpful for knowing the attitude of a learner, and useful for evaluating a learner.

We take the same approach for the estimation of those three internal states. The ground truth values for all three internal states need to be gathered in terms of introspection, because the internal states cannot be physically measured by current technology in actual e-learning environments. For their estimation, we choose visual sensing, because a non-intrusive way without heavy constraints and cost is preferable.

### 4.2 Scoring of Learning States

Some previous studies used evaluation by trained judges as ground truth ([9], [17], [18]). However, such judgments are often different from self-evaluation, and consequently, the ground truth indicates how learners look more than how they feel. In contrast, we use learners' self-evaluation as ground truth. To avoid learning interruptions and obtaining high reliance on self-reports, we integrate several methods summarized in [19] by comparing their advantages and limitations. The details regarding ground truth acquisition will be described in Sect. 7. However, self-reporting introduces the problem of a tendency of participants to average their ratings [19], as well as the issue of social desirability.

In most previous work, two-level evaluation was used, e.g. attentive vs. inattentive or boredom vs. neutral, as was suitable for their purposes. It is often difficult for humans to evaluate themselves quantitatively. However, multiple-level measurement is commonly used in psychometrics to measure attitudes for analysis, e.g., the Likert scale or semantic differentials. For the benefit of teachers, we use a five-level scale for learning state measurement as shown in Table 1. Nevertheless, this introduces ambiguity and differences among persons. This problem will be discussed in the following sections.

## 5. Learning State Recognition by Visual Sensing

Each learning period is segmented into intervals of a specified length, which is 30 seconds in our experiments. For each interval, the following features are detected, and the learning states are estimated.

### 5.1 Low-Level Feature Detection

Low-level features, including three-dimensional (3D) head pose (position and angle), facial parts movements, and body area and distance, are obtained from RGB-D images. Head pose is obtained using face detection, and movements of mouth and eyebrows are obtained using facial parts detection. These detections are based on an active appearance model [20]. Head pose is indicated in terms of translation and rotation angles in camera coordinates. Movements of mouth and eyebrows are indicated by displacements from the neutral position and shape of mouth and eyebrows.

### 5.2 Intermediate Feature Detection

Three categories of intermediate feature are obtained using low-level features.
(1) Presence information: Presence of the learner in front of the screen and the distance between the learner and the screen. We use face detection and body area measurement results for this purpose. Specifically, when the algorithm fails to detect the face, and the body size is smaller than the threshold, a learner is considered as "absent." Otherwise, the learner is "present," and the distance is obtained from the depth in the RGB-D image.
(2) Head and facial parts information: These features are obtained directly from the head pose and facial parts movements.
(3) Probability of gazing at the screen: We do not use commercial eye-gaze tracking systems because of the cost. Alternatively, gazing direction, i.e., a line of sight, is estimated by face orientation based on training samples. If we assume that a learner is looking straight forward, then gazing direction is the same as face orientation, and the gazing target is the "intersection" of the line of sight and the environment. However, this does not always hold. To cope with this problem, we use the statistics of samples collected beforehand, in which a participant looked "inside" and "outside" of the screen with changing directions and conditions. The monitor screen area is quantized into 10 x 10 small regions called cells. The probability of "gazing at the monitor screen" for each cell is calculated based on Bayes' theorem, using the following formula:

$$P(A|X_i) = \frac{P(X_i|A)P(A)}{P(X_i)} \tag{1}$$

where $A$ indicates that a participant looks inside the screen, and $X_i$ indicates that the participant's line of sight intersects the $i^{th}$ cell.

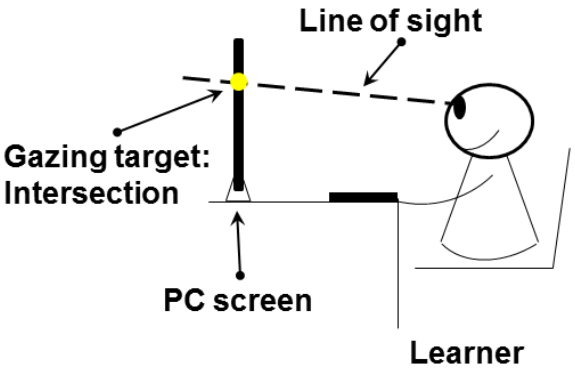


**Fig. 2** Gazing target: intersection of line of sight and screen

**Table 2** Thirty-three-element feature vector

| Feature source: | Feature items: |
|---|---|
| Presence information | Present proportion |
| | Distance (Max, Average, Min) |
| Head and facial parts information | Face detection successful proportion |
| | Lips movement (Max, Average, Min) |
| | Eyebrows movement (Max, Average, Min) |
| | Head pose angle Pitch (Max, Average, Min) |
| | Head pose angle Yaw (Max, Average, Min) |
| | Head pose angle Roll (Max, Average, Min) |
| | Head position X-coordinate (Max, Average, Min) |
| | Head position Y-coordinate (Max, Average, Min) |
| | Head position Z-coordinate (Max, Average, Min) |
| Probability of gazing at screen | High probability proportion |
| | Moderate probability proportion |
| | Low probability proportion |
| | Consider as zero probability proportion |

### 5.3 Classification by Support Vector Machine

A feature vector with 33 elements is calculated based on intermediate features of each interval, as listed in Table 2. Together with the self-evaluation score as ground truth, they comprise one training sample.

We use an SVM for classification, specifically, the one-against-one method for handling multiclass classification in LIBSVM [21]. A radial basis function is used as the kernel. Details on our experimental data will be provided in Sect. 7.

## 6. Inter-personal Differences

### 6.1 Strategy for Adjusting to a New Learner

For learners to record their learning states after actual e-learning requires considerable effort. Scoring, often takes time longer than the time for actual learning, and requires considerable mental effort in video reviewing, introspection, and marking†. Provided scores by learners would be unreliable due to the additional load, if learners are simply forced to mark scores after actual e-learning. On the other hand, during the phase of system development, we can expect that a considerable number of samples can be obtained from multiple participants. From this point of view, we need a system that can adjust quickly to a new learner with little effort by the learner. For this purpose, we consider the following scheme:

Step 1: We gather sufficiently many samples from multiple learners who collaborate for data collection. Hereafter, we call those learners and samples "prototype learners" and "prototype samples," respectively.

Step 2: The system asks a new learner to provide ground truth scores for a small number of intervals in actual e-learning. Hereafter, we call these "representative samples."

Step 3: The system selects appropriate classifiers by using both representative and prototype samples.

†Details of ground truth acquisition will be described in 7.1.

### 6.2 Classifier Selection Strategies

Assuming that we have sufficiently many prototype samples from multiple prototype learners, we can think of various classifiers as follows:

(C1) Classifiers that are each trained with prototype samples from a single prototype learner.

(C2) Classifiers that are each trained with prototype samples from a combination of two or more prototype learners.

(C3) A classifier that is trained with all prototype samples from all prototype learners. Hereafter, we call this the "unified classifier."

The possible variations comprise the power set of the learners. With those classifiers, our target is to develop a method for choosing the best classifier for a new learner. We consider the following strategies for this problem:

(M1) Accuracy-based method: The system applies every classifier to the representative samples from a new learner and then chooses the classifier that results in the best performance.

(M2) Similarity-based method: The system measures the similarities between representative samples of a new student and prototype samples and then chooses the classifier that has the greatest number of similar samples in its training data within a specified number of resembling samples. In this approach, L1 distance is chosen as the metric for similarity measurement. This strategy has the advantage that it does not require self-evaluation by a new learner.

## 7. Experimental Results

### 7.1 Environment of Experiment

As e-learning contents, we chose multimedia English learning materials for the Computer-Assisted Language Learning (CALL) system at Kyoto University [22]. These materials are used widely by students not only at Kyoto University but also in several other universities. Learners are university students expected to have self-regulatory e-learning. A sample image is shown in Fig. 3.

A Kinect camera is located above the monitor for recording learners' video and estimating internal states. The computer screen is captured continuously using a frame

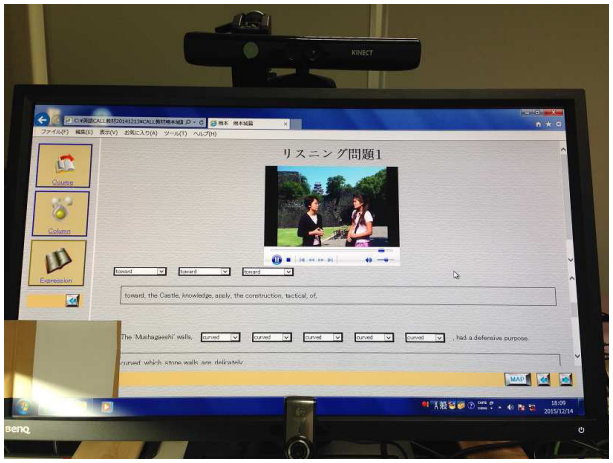


**Fig. 3** E-learning interface

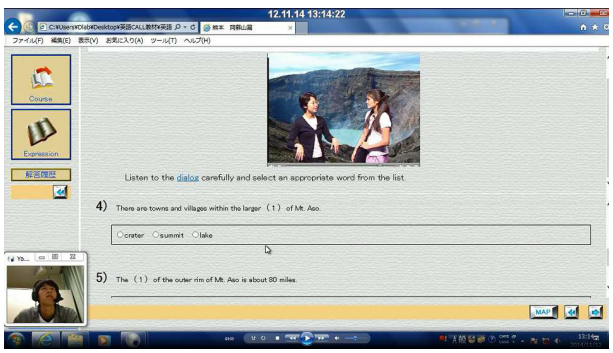


**Fig. 4** Reviewing for self-scoring

grabber. A small webcam is placed at the foot of the monitor to capture the learner's face independently of the Kinect camera. Both the learner's face and the screen capture can be shown synchronously to the learner right after the e-learning session, as shown in Fig. 4, and the prototype learner provides self-evaluation for each interval by watching them.

We gathered seven participants, undergraduate students with no experience in learning with the specified e-learning materials. Each participant participated in an average of nine sessions, each of which was approximately 30 minutes in length. We obtained samples for 1,559.5 minutes, i.e., 3,119 valid samples of intervals with self-evaluation scores as ground truth.

### 7.2 Potential of Learning States Estimation

We first examined the potential performance of simple classification accuracy by the following schema:
(E1) A classifier is trained using samples from one participant mentioned as C1 in Sect. 6.2. Target samples, i.e., recognition targets, are from the same participant.
(E2) A classifier is trained using samples from one participant mentioned as C1. Target samples are taken from a different participant.
(E3) A classifier is trained using samples from all participants mentioned as C3, excluding samples from the participant considered as a new learner. Target samples are from the excluded participant.

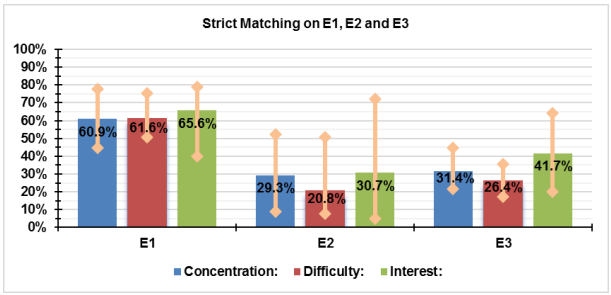


**Fig. 5** Strict matching performance on E1, E2, and E3 conditions.

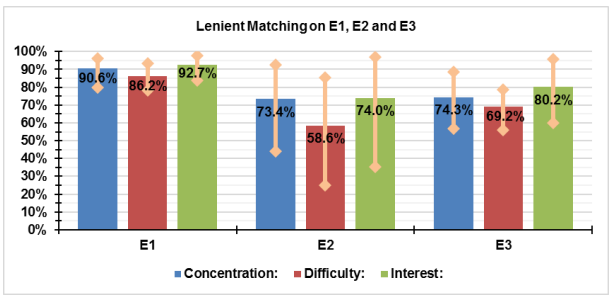


**Fig. 6** Lenient matching performance on E1, E2, and E3 conditions.

E1 estimates the upper bound performance of our method. In other words, the performance we can expect if the learner is well-known. E2 shows the performance degradation caused by inter-personal differences. E3 estimates the potential performance of the unified classifier for a new learner. For each scheme, the leave-one-out method is used for cross validation.

The criteria of the matching are as follows:
1. Strict matching criterion: This requires an exact match between the classification result provided by the SVM and the ground truth.
2. Lenient matching criterion: This allows a classification result to be equal or nearly equal to the ground truth, i.e., the score difference must be at most one.

Lenient matching criterion is introduced to examine how estimated scores are close to the ground truth. If the performance of lenient matching is satisfactory, an estimated score can be useful even if it does not perfectly match to the ground truth. Another aspect of lenient matching is concerning difficulty of introspection. Because introspection may contain fluctuation caused by human nature, strict matching criterion could be too severe, and teachers may feel that lenient matching criterion is reasonable.

Figures 5 and 6 show the results for the strict and lenient matching criteria, respectively. The average values are shown by thick bars, and the ranges between the best case and the worst case are shown as thin lines.

From the E1 results, we can see that the average accuracy of strict matching is approximately 60%. One of

**Fig. 7** Similar appearances but different self-evaluation scores of the same learner. Top sample: Reference sample, self-evaluation score 1. Bottom-left sample: Closest sample 1, self-evaluation score 3. Bottom-right sample: Closest sample 2, self-evaluation score 4

**Fig. 8** Different situations for same self-evaluation score: score 1 in concentration-distraction

the primary reasons for the performance degradation in this case is the similarity of the behaviors among different internal states. Figure 7 shows three samples with different self-evaluation scores for concentration-distraction. The first sample is chosen as the reference, for which the L1 distances to the other samples of the same student are calculated. The other two are the closest samples. During these three intervals, the learner retained appearances as in the shown images, with only small movements. Consequently, the feature vectors for these intervals are similar, while the self-evaluation scores are different.

Another reason is the dissimilarity of behaviors for the same self-evaluation score. An example is shown in Fig. 8. Four representative images result in four different situations with score 1 for concentration-distraction producing diverse feature vectors. The top-left image is the same sample as the top sample in Fig. 7. The top-right sample indicates that the learner is napping; the bottom-left sample indicates that the learner is almost out of field. In the bottom-right case, the learner does not look at the screen. The feature vectors for those three samples are much different from those for the first sample in the L1 distance metric.

Another possible reason is the ambiguity of self-evaluation. Self-evaluation appears to fluctuate because of the difficulty of introspecting. From this point of view, we cannot expect perfect performance. However, all of those drawbacks are relaxed in the lenient matching cases. We obtain approximately 90% accuracy for all three pairs of internal states.

As the results of E2, we have serious performance degradation for strict matching if we apply classifiers trained for a different person, as shown in Fig. 5. Figure 9 shows an example of similar appearances for different learners. The first image is the reference sample with score 5 for concentration-distraction, for which L1 distances to the other students' samples are calculated. The two closest sam-

**Fig. 9** Similar appearances but different self-evaluation scores of different learners. Top sample: Reference sample, self-evaluation score 5. Bottom-left sample: Closest sample 1, self-evaluation score 3. Bottom-right sample: Closest sample 2, self-evaluation score 1.

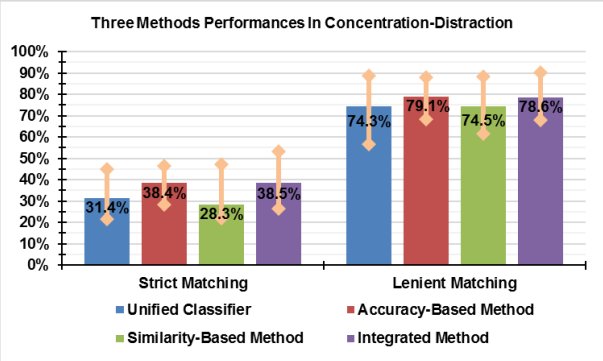


**Fig. 10** Three methods for performances in concentration-distraction

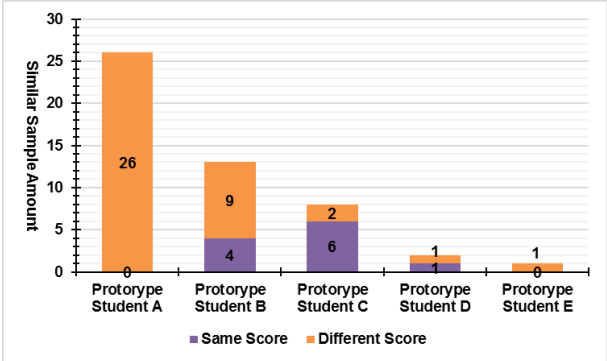


**Fig. 11** Score distribution of similar samples

ples, whose self-evaluation scores are 3 and 1, respectively, are shown in Fig. 9. Based on the images, they are not significantly different except in the detailed conditions of the eyes that are not included in feature vectors in our experiments. Detection of those detailed features is left for future work. The performance degradation is, however, relaxed with the lenient matching criterion, as shown in Fig. 6.

The E3 results show the performance expectation of the unified classifier for a new learner. The large gaps between E1 and E3 indicate that simple aggregation of a large number of samples does not necessarily provide a good classifier. On the other hand, the performance difference between E3 and E2 shows that the performance improvement resulting from gathering samples from multiple persons is not negligible. Those results imply that we need a more sophisticated method to utilize a variety of samples from a variety of learners effectively. This problem is discussed in 7.4.

Concerning the differences among learning states, performance is not significantly different. However, we observe the tendency that the accuracy is better in the order of interest-boredom, concentration-distraction, and difficulty-ease. One possible reason might be a negative mood that learners choose to conceal.

### 7.3 Performance for a New Learner

Experiments were conducted for checking the possibility of choosing an appropriate classifier by using a small number of representative samples from a new learner. For this purpose, one participant was chosen as a new learner in turn, and the other participants were regarded as prototype learners. Five representative samples were chosen randomly from the new learner's samples. We applied the accuracy- and similarity-based methods for choosing a classifier from among all classifiers, i.e., C1, C2, and C3 in 6.2. Then, the chosen classifier was applied to all the samples of the new learner, and the performance was evaluated. We repeated this process 20 times for every new learner, and the average performance was recorded.

Figure 10 shows the results for concentration-distraction. The baseline is the performance by the unified classifier. The average accuracy of seven new learners is displayed with a thick bar. The values of the highest and the lowest accuracy are presented by the thin lines on the bars. We can see that the accuracy-based method provides better performance than the unified classifiers. On the other hand, the similarity-based method does not provide good results. Figure 11 shows one example that might explain this. For each of five representative samples of new learners, the ten closest samples are extracted from all prototype samples, and their scores are counted. Although prototype learner A has the most samples similar to representative samples of the new learner, none of them has the same score. This fact clearly shows the difficulty posed by inter-personal differences.

Next, we examined the integrated method, a combination of the above two methods. The first process is the same as the accuracy-based method. If there are multiple classifiers that yield the best performance for given representative samples, the second process chooses the classifier that has the greatest number of similar samples out of the 50 closest samples. As shown in Fig. 10, the integrated method shows no significant improvement over the accuracy-based method. This might be a result of the small number of representative samples, in addition to the facts disclosed in Fig. 11. The results for difficulty-ease and interest-boredom are shown in Figs. 12 and 13, respectively. They also indicate no significant improvements from the integrated method.

As for difficulty-ease, the results in E2 and E3 that have the smallest average accuracies and the smallest range between the highest and the lowest accuracy suggest larger inter-personal differences among learners. This diversity makes improvements difficult. For the interest-boredom case, the performance of the baseline, i.e., E3, is better than that of the other two learning states. The room for improvement is consequently small.

### 7.4 Discussion

The results of E1 show much room for improvement, especially in strict matching. One reason is the wide variety of external expressions, and another is the ambiguity of self-evaluation. For future improvements, incorporating other

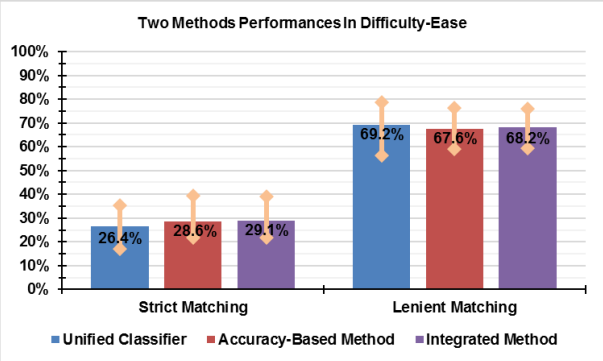


**Fig. 12** Two methods for performances in difficulty-ease

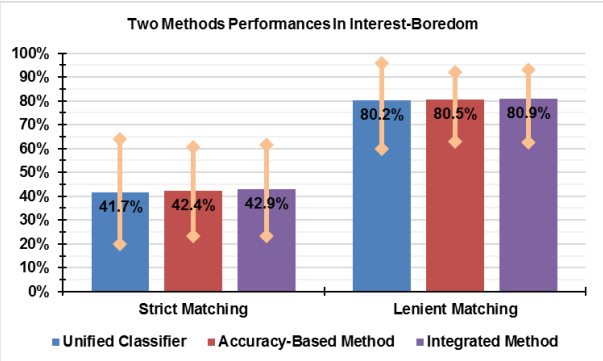


**Fig. 13** Two methods for performances in interest-boredom

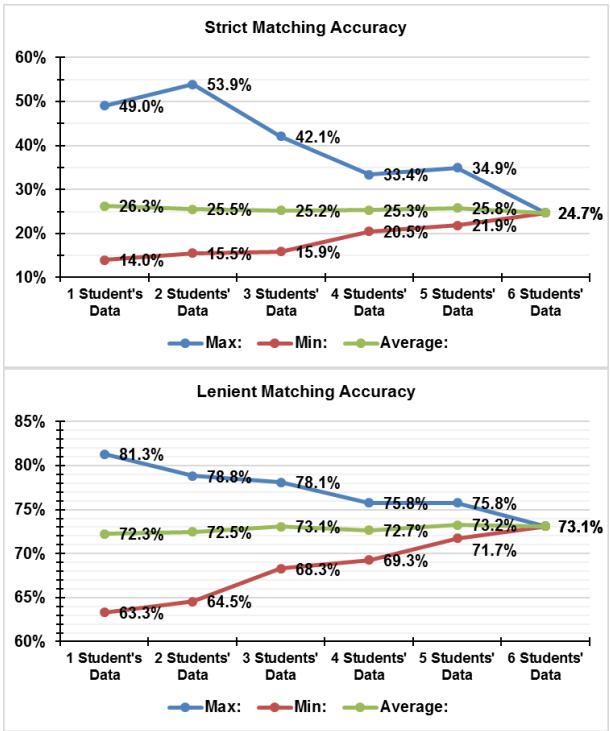


**Fig. 14** Performance changes by the number of prototype learners

measuring modalities might be a partial solution. Detailed information regarding the eyes is expected to improve the performance. Additionally, we can consider the use of non-intrusive physiological measurements and/or the use of a mouse and keyboard to input information.

E2 results show serious inter-personal differences, and E3 results reveal that simply mixing all samples does not provide sufficient improvement. To clarify this point, we verified how performance changes with the number of prototype learners. Figure 14 shows the result. The horizontal axis indicates the number of prototype learners used for training, and the vertical axis indicates the accuracy of estimating concentration-distraction for a new learner. The three lines show maximum accuracy, average, and minimum accuracy.

The figure illustrates the trend that the maximum accuracy is better with one or a few prototype learners, and it worsens as the number of prototype learners increases. This fact suggests that a classifier trained by the samples from one or a few similar prototype learners tends to provide good performance. On the other hand, the minimum accuracy improves gradually as the number of prototype learner increases. A classifier trained for one or a few prototype learners with different characteristics yields poor performance, and this problem will be relaxed with more prototype learners. As the number of prototype learners increases, the chance that samples with behavioral characteristics similar to a new learner's are included will increase. This fact suggests that it is useful to consider the unified classifier based on all of the samples as the baseline.

Next, we focus on the selection of an appropriate classifier for a new learner. In our research, we use only a small number of representative samples, viz., five samples in our experiments, in order to keep the workload of a new learner as low as possible. The accuracy-based and the integrated methods improved the estimation accuracy in some cases, but did not provide significant improvements in other cases. From the results of E1 and Fig. 14, we see that there are many classifiers that provide better performance. We have much room for improvement for future work. We can expect better accuracy using the integrated method if we can obtain more representative samples with less effort by a new learner. Therefore, a method for reducing effort in and distress over self-scoring can be expected to lead to accuracy improvements.

## 8. Conclusion

In this research, we designed an e-learning support system that can capture learners' behaviors visually and estimate learners' learning states. We chose concentration-distraction, difficulty-ease, and interest-boredom as a learner's learning states, and these were recognized by using the learner's presence, head position, and facial feature information. The experimental results showed the potential of our classification method by SVM using the abovementioned visual features: approximately 60% average accuracy in strict matching and approximately 90% average accuracy

in lenient matching can be achieved. We also examined practical methods for adjusting to a new learner who can provide only a few samples as ground truth. Accuracy-based selection of classifiers and our integrated method showed better performance than the unified classifier for which all of the samples were used for classifier training.

For future work, we need a variety of investigations to improve recognition accuracy, including improvements in the sensing system, feature selection, and classifier selection. We also need experiments with various e-learning materials and learners of diverse ages, as well as practical use in classes. Developing a user interface for providing educational information for teachers' browsing is also important.

**References**

[1] N.J. Butko, G. Theocharous, M. Philipose, and J.R. Movellan, "Automated facial affect analysis for one-on-one tutoring applications," Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp.382–387, IEEE, 2011.
[2] M.B. Ammar, M. Neji, A.M. Alimi, and G. Gouardères, "The affective tutoring system," Expert Systems with Applications, vol.37, no.4, pp.3013–3023, 2010.
[3] J. Whitehill, M. Bartlett, and J. Movellan, "Automatic facial expression recognition for intelligent tutoring systems," Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pp.1–6, IEEE, 2008.
[4] K. Zakharov, A. Mitrovic, and L. Johnston, "Towards emotionally-intelligent pedagogical agents," in Intelligent Tutoring Systems, pp.19–28, Springer, 2008.
[5] S.K. D'Mello, S.D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser, "Automatic detection of learner's affect from conversational cues," User modeling and user-adapted interaction, vol.18, no.1-2, pp.45–80, 2008.
[6] D. Litman and K. Forbes, "Recognizing emotions from student speech in tutoring dialogues," Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on, pp.25–30, IEEE, 2003.
[7] L. Shen, M. Wang, and R. Shen, "Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment," Educational Technology & Society, vol.12, no.2, pp.176–189, 2009.
[8] C.-H. Yang, "Fuzzy fusion for attending and responding assessment system of affective teaching goals in distance learning," Expert Systems with Applications, vol.39, no.3, pp.2501–2508, 2012.
[9] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard, "Affect-aware tutors: recognising and responding to student affect," International Journal of Learning Technology, vol.4, no.3-4, pp.129–164, 2009.
[10] J.A. Fredricks, P.C. Blumenfeld, and A.H. Paris, "School engagement: Potential of the concept, state of the evidence," Review of educational research, vol.74, no.1, pp.59–109, 2004.
[11] D.J. Shernoff, "The nature of engagement in schools," Optimal Learning Environments to Promote Student Engagement, pp.47–75, Springer, 2013.
[12] M. Csikszentmihalyi, Flow. the psychology of optimal experience, Harper Perennial, New York, 1990.
[13] E.L. Deci, "The relation of interest to the motivation of behavior: A self-determination theory perspective," The role of interest in learning and development, pp.43–70, 1992.
[14] U. Schiefele, A. Krapp, and A. Winteler, "Interest as a predictor of academic achievement: A meta-analysis of research," The role of interest in learning and development, pp.183–212, 1992.
[15] L.L. Shirey, "Importance, interest, and selective attention," The Role of Interest in Learning and Development, pp.281–296, 1992.
[16] P.J. Silvia, "Self-efficacy and interest: Experimental studies of optimal incompetence," Journal of Vocational Behavior, vol.62, no.2, pp.237–249, 2003.
[17] S. D'Mello, R.W. Picard, and A. Graesser, "Toward an affect-sensitive autotutor," IEEE Intelligent Systems, vol.22, no.4, pp.53–61, 2007.
[18] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," Proc. 29th Annual Cognitive Science Society, pp.467–472, CiteSeer, 2007.
[19] M. Wosnitza and S. Volet, "Origin, direction and impact of emotions in social online learning," Learning and instruction, vol.15, no.5, pp.449–464, 2005.
[20] N. Smolyanskiy, C. Huitema, L. Liang, and S.E. Anderson, "Real-time 3d face tracking based on active appearance model constrained by depth data," Image and Vision Computing, vol.32, no.11, pp.860–869, 2014.
[21] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol.2, no.3, p.27, 2011.
[22] "Kyoto university call learning space," http://www.kyoto-u.ac.jp/ja/education-campus/curriculum/foreign/call.html

**Siyang Yu** received the B.S. and M.S degree in Educational Technology from Beihua University and Northeast Normal University, China, in 2006 and 2010. He is currently working toward the Ph.D. degree in Graduate school of Engineering, Kyoto University. His current research interests include e-learning supportive system and intelligent CAI.

**Kazuaki Kondo** received his M. E. and Ph. D. degrees from Osaka University in Japan. He became a research associate at Osaka University in 2007, an assistant professor at Kyoto university in 2009, and a lecturer in 2015. He was awarded the Kusumoto award in 2002. His research interests are computer vision and intelligent support on human communications. He is a member of IEICE.

**Yuichi Nakamura** received B.E, M.E, and Ph.D degrees in electrical engineering from Kyoto University, in 1985, 1987, and 1992, respectively. From 1990 to 1993, he worked as an instructor at the Department of Electrical Engineering of Kyoto University. From 1993 to 2004, he worked for Institute of Information Sciences and Electronics of University of Tsukuba, Institute of Engineering Mechanics and Systems of University of Tsukuba, as an assistant professor and an associate professor, respectively. Since 2004, he has been a professor of Academic Center of Computing and Media Studies, Kyoto University. His research interests are on computer vision, multimedia, human-computer and human-human interaction including distance communication, and multimedia contents production.

**Takayuki Nakajima** received the B.S. a degree in Human Institute and M.S. a degrees in Human Environmental Studies from Kyoto University in 2012 and 2014, respectively. Since 2014, he has managed in self access center for language learning under Professor Dantsuji.

**Masatake Dantsuji** received the B.S. and M.S. degrees in Letters from Kyoto University in 1979 and 1981, respectively.During 1990-1997, he stayed in Kansai University as associate professor. From 1997, he is a professor of Kyoto University.