Estimating Work Situations from Videos of Practical Training Classes with Assembly Tasks

K. Okamoto, K. Kakusho, M. Yamamoto, T. Kojima, and M. Murakami

***Abstract*—Our preceding study proposed the possibility of producing video previews that enhance viewer motivations towards assembly work in practical training classes by picking scenes of work situations with good performance from the videos of past classes. In the study, two conceptual attributes for categorizing work situations from the viewpoint of the performance were introduced with reference to previous studies for evaluating productivity of human intellectual work with computers. Based on those two conceptual attributes, our preceding study employed observable features for estimating work situations in the videos and showed that those features seem to reflect the difference of work situations with respect to the conceptual attributes. However, quantitative precision for estimating work situations from those features has not yet been evaluated. Moreover, those observable features are employed without considering whether humans actually pay attention to them. It is also not clear whether videos with work situations sufficient for each of the conceptual attributes actually enhance viewer motivations towards the work. This article clarifies these issues based on our recent experimental results with experimental participants.**

***Index Terms*—Practical training class, assembly task, work situation estimation, observable feature, physical activity, mental concentration.**

I. Introduction

With recent progress in the introduction of information technologies to the field of education, it has become typical for various educational institutions, especially universities and colleges, to take videos of actual classes. One of the typical uses of these videos is viewer learning, such as massive open online courses (MOOCs) [1]. Another use is evaluation of the classes by the instructors to facilitate further improvement.

Conventionally, videos of actual classes are mainly taken for mass classroom lectures where all the attendees listen to the talks by the lecturers to acquire knowledge of specific areas. However, it is also useful to take videos of practical training classes where each trainee performs practical work individually to acquire a particular skill. Those videos are not only useful for instructors to grasp the work situations of each trainee for further improvement of instruction methods but also for producing video previews that enhance motivations of new trainees towards the practical work before they start to work themselves. Actually, a previous study of *self-modeling* [2] claimed that repetitive viewing of videos including successful actions for accomplishing a task enhances the confidence and motivation of the viewer, and it has been shown that learning by watching those videos is effective for increasing the *self-efficacy* [3] and performance of the learners in educational institutions including universities [4]–[6].

Our preceding study [7] proposed a possibility of producing video previews to enhance motivation of trainees towards practical work by picking the scenes of work situations with good performance from the videos of past classes. Since it is necessary to realize a kind of video indexing process that estimates work situations of each trainee in the videos to produce such video previews automatically, our preceding study discussed the possibility by focusing on an assembly task as a representative example of the work for practical training classes. In the study, the physical activity involved in the assembly work and the mental concentration of the trainees on the assembly work were introduced as conceptual attributes to categorize work situations, because those attributes are also considered in previous studies evaluating productivity of human intellectual work with computers [8]. For the observable features that are useful to discriminate the different situations with respect to those two attributes and obtainable from videos by image processing, our preceding study employed the distance from the face of each trainee to work objects and the temporal change in the dispersion of their positions on the work surface. From the experimental result of the study, it appears that the values of those observable features take different values for the video clips with different work situations in those attributes.

However, the preceding study described above does not quantitatively evaluate precision for estimating work situations with respect to the two conceptual attributes from the employed observable features. Moreover, in the first place, the study employs those observable features without considering whether humans actually paid attention to them. It is also not clear whether our motivations towards the assembly work are actually enhanced by watching videos with work situations estimated to be sufficient with respect to the two conceptual attributes. In summary, the following unanswered questions remain:

1) What observable features do humans actually pay attention to when we categorize work situations with respect to physical activity and mental concentration? Are those observable features similar to those employed in our preceding study?
2) How precisely can the work situations be estimated with respect to the two conceptual attributes from the

Manuscript received December 30, 2016; revised February 3, 2017.

Kai Okamoto, Koh Kakusho, and Michiya Yamamoto are with School of Science and Technology, Kwansei Gakuin University, Sanda, Japan (e-mail: {kaisea, kakusho, michiya.yamamoto}@kwansei.ac.jp).

Takatsugu Kojima is with Faculty of Medicine, Shiga University of Medical Science, Otsu, Japan (e-mail: kojima@kojima-lab.net).

Masayuki Murakami is with Research Center for Multi-Media Education, Kyoto University of Foreign Studies, Kyoto, Japan (e-mail: masayuki@murakami-lab.org).

doi: 10.18178/ijiet.2018.8.1.1009

observable features in (1)?

3) Are videos of assembly work with different work situations estimated from the observable features in (1) with the precision in (2) different for enhancing the motivations of the viewers towards the work?

In this article, these questions are resolved based on our recent experimental results with experimental participants.

A brief introduction of the conceptual attributes as well as observable features employed in our preceding study is given in Section II. In Section III, question 1) is discussed based on our experimental results with experimental participants. Based on the experimental results, some candidate observable features are considered, and the best observable feature is determined for each of the two conceptual attributes in Sections IV and V. In Section VI, precision of the work situations estimated from the best observable features is evaluated to answer question 2). Motivations of the viewers who watch videos with different work situations towards the assembly work are also evaluated to resolve question 3) in this section. Concluding remarks and future work are described in Section VII.

II. Our Preceding Study

A. Conceptual Attributes for Categorizing Situations of Assembly Work

In many previous studies on the methodology for evaluating productivity of human intellectual work mainly with computers, the productivity is evaluated primarily with respect to the physical activity involved in the work in progress and the worker's mental concentration on the work. Some of these studies focus on either of the two conceptual attributes [9]–[15] while the others consider both [8]. Since these two attributes are not specific for particular tasks but, rather, are general portions applicable to any physical work by humans having mental states, our preceding study introduced the same two attributes for categorizing assembly work situations of practical training where the physical activity means how actively each trainee performs the required assembly work and mental concentration means how closely the attention of the trainee is concentrated on the assembly work. As a result, work situations are classified into four categories: *high activity & high concentration*, *high activity & low concentration*, *low activity & low concentration*, and *low activity & high concentration*.

B. Observable Features Reflecting Difference in Work Situations

To estimate work situations from observation of actual work, observable features useful for discriminating differences in work situations need to be employed. Previous studies for evaluating human intellectual work with computers employ amount and types of operations performed on the computers as observable features for evaluating physical activity of the work, because every operation performed on the computers can be easily obtained from their log data. However, in assembly work performed in the physical world, it is not easy to recognize the type of each operation by image processing for the videos of the work. Thus, our preceding study simply measures the temporal change in the dispersion of the positions of the work objects scattered on the work surface as the observable feature of the physical activity.

For an observable feature of mental concentration, previous studies for human intellectual work evaluation employ the distance from the face of the worker to the display of the computer. By referring to these studies, our preceding study employs the distance from the face of each trainee to the work surface as the observable feature of mental concentration on assembly work. Both of these observable features of the physical activity and mental concentration for assembly work are obtained from video images by image processing as described in the next section.

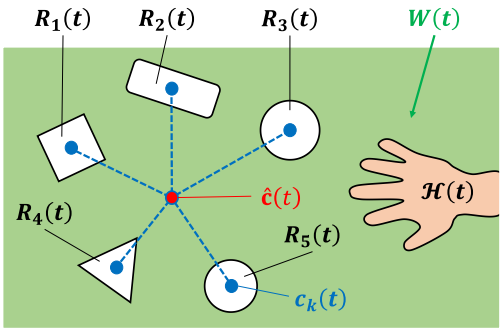


Fig. 1. Image features employed to evaluate the dispersion for the position of work objects on the work surface [7].

C. Difference in Observable Features Obtained from Videos of Assembly Work Situations

As described in *B*, the observable feature of the physical activity of assembly work is given by temporal change in the dispersion for the positions of work objects in our previous study. This observable feature is obtained by image processing for the images captured by a camera installed right above the work surface. This camera is called the *overhead camera* hereafter. All the pixels of image frame at any moment *t* are separated into the work surface region and the hand region denoted by $W(t)$ and $\mathcal{H}(t)$, respectively, as well as work object regions denoted by $R_1(t), R_2(t), \cdots$, based on their colors. The dispersion for the positions of work objects at *t* is denoted here by $d(t)$, which is calculated as the sum of squared deviations of the work object region centroids denoted by $\boldsymbol{c}_1(t), \boldsymbol{c}_2(t), \cdots$, from their average position denoted by $\hat{\boldsymbol{c}}(t)$ (see Fig. 1). The temporal change in the dispersion during any period $[t, t+\Delta t]$ in a video is denoted by $D[t, t+\Delta t]$, which is defined as follows:

$$D[t, t+\Delta t] = \sum_{\tau=t}^{t+\Delta t-1} |d(\tau+1) - d(\tau)|$$

For the observable feature of mental concentration, the distance between the face of any trainee and the work surface needs to be calculated in our previous study, as described in *B*. This distance is obtained by facial image processing for the images captured by another camera installed in front of the trainee. This camera is called the *front camera* hereafter. For simplicity, the distance from the face to the work surface is approximated as that from the face to the front camera along with its optical axis. The three-dimensional (3D) position of

the face in the front-camera-centered coordinate system is denoted by $\boldsymbol{f}^C(t)$, which is estimated from the size and the orientation of the face obtained by facial image processing for the image frame of the front camera at t. The approximated distance from the face to the front camera is given as z coordinate of $\boldsymbol{f}^C(t)$. The average of this z coordinate during any period $[t, t+\Delta t]$ is employed as the observable feature of mental concentration.

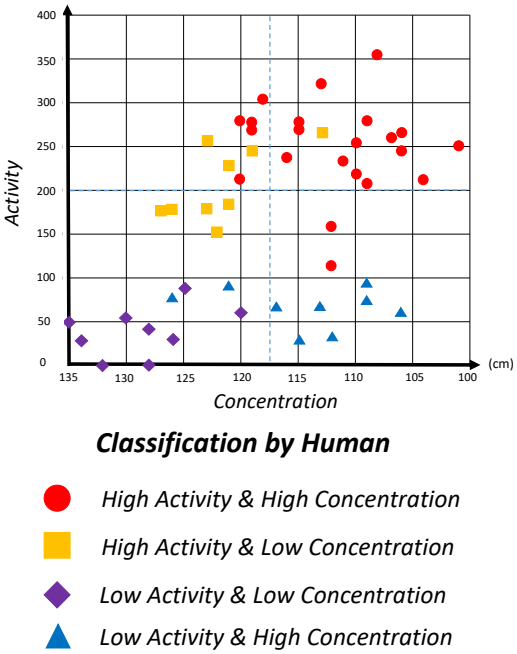


Fig. 2. Measurements of the observable features for the physical activity and mental concentration for video clips manually classified into different categories of work situations [7].

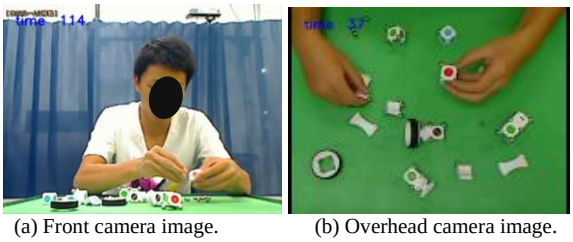
In the experiment in our preceding work, an assembly task using a *moss robot* [16] is performed by six experimental participants who play roles of trainees in a practical training class. The symbols on the graph of Fig. 2 illustrate pairs of the values for the two observable features for 50 video clips of various work situations by their positions on the graph. Each video clip is 10 seconds long and includes a pair of images captured by the overhead and front cameras. Work situations in those video clips are manually classified into the four categories described in section *A* and displayed by different symbols in the graph. As shown in this figure, the video clip classified as the same category seems to take similar values for the observable features.

The above result of our preceding study implies that the observable features roughly reflect the difference in work situations with respect to the two conceptual attributes. However, it is not quantitatively evaluated how precisely those observable features can actually estimate the correct categories of work situations. Moreover, in the first place, it is not discussed what our humans actually pay attention to as observable features for discriminating different work situations. It is also not evaluated whether the videos with the work situation estimated as *high activity & high concentration* actually enhance the motivation of viewers towards the assembly work more than the other videos. In the following sections, these issues are discussed based on evaluation of the videos with many experimental participants.

III. Finding Observable Features to Which Humans Pay Attention

A. Experimental Settings

The objective of the first experiment in this article is to clarify what our humans actually pay attention to when we evaluate the physical activity and mental concentration by watching videos. In this experiment, 151 experimental participants were asked to categorize work situations appearing in video clips. These video clips are the same as those employed in our preceding study described in section II.*C*. However, the facial region in each image frame of the front camera is masked in this experiment so that the facial expression does not affect evaluation by experimental participants. This is because human facial expressions could be interpreted in various ways depending on the viewer's subjective viewpoints, and this article only considers observable features in general viewpoint for the first step. Sample images are shown in Fig. 3.



(a) Front camera image. (b) Overhead camera image.

Fig. 3. Sample images in video clips obtained by the front and overhead cameras. The facial region is masked as shown in (a).

The 50 video clips obtained above are divided into five groups, each of which includes 10 video clips. Three groups among them are presented to 83 experimental participants, and two groups are presented to 68 participants. Each participant is asked to classify the work situation in each video clip included in the presented groups into one of the four categories described in section II.*A* and to describe the reason why the video clip is classified into such category, using the following procedure:

1) Receive an explanation about the meaning of the physical activity and mental concentration with an example of Japanese *origami* (paper folding), which has no relevance to the assembly work to be evaluated later.
2) Categorize the physical activity and mental concentration of the work situation in the presented video clip as *high* or *low*.
3) Give the reason for the categorization in (2) by free descriptive texts.
4) Repeat steps (2) and (3) for all the video clips included in the presented groups.

B. Result of Morpheme Analysis

To find the observable features that humans actually pay attention for estimating work situations with respect to the physical activity and mental concentration, a morpheme analysis is provided for the free descriptive texts in the answers obtained in step (3). Major words frequently used to describe the reason for categorizing work situations as *high* or *low* for the physical activity and mental concentration are extracted by the analysis. Table I shows English translation of those major words in descending order of the number of occurrences.

TABLE I: OCCURRENCE OF MAJOR KEY WORDS IN THE ANSWERS FOR THE REASON FOR CLASSIFYING THE WORK SITUATIONS

(a) With respect to the physical activity

High		*Low*	
"hand"	226	"not"	269
"parts"	126	"hand"	234
"move"	122	"move"	182
"fast"	39	"parts"	106
"task"	37	"hold"	93
"motion"	36	"stop"	41
"assemble"	28	"little"	24

(b) With respect to the mental concentration

High		*Low*	
"parts"	165	"not"	175
"gaze"	126	"parts"	111
"forward"	66	"gaze"	72
"face"	50	"back"	57
"think"	46	"hand"	42
"task"	42	"task"	40
"posture"	24	"face"	29

As shown in the table, negative words, translated here as "not," are used much more frequently in the description of the reason for categorizing work situations as *low* than that for categorizing them as *high*, regardless of the conceptual attributes. Since frequent use of negative words for the categories of *low* implies disappearance of certain observable features in those situations, frequently used words for situations categorized as *high* should be considered as possible observable features.

The description of the reason for giving the category of *high* for the physical activity often includes words meaning "hand" or "parts" as well as those meaning "move." Since the words meaning "fast," "task," "motion," and "assemble" also occur with a certain frequency, humans seem to pay attention to the amount or speed of the motion of the hands or work objects as well as the progress of the assembly task to categorize the work situations of the physical activity.

For mental concentration, on the other hand, words meaning "parts" and "gaze" occur relatively more frequently than the others. Words meaning "forward" and "face" also appear with a certain frequency. From these words, the direction or the distance from the face to work objects seems to be attended to.

Based on the result of the morpheme analysis above, the observable features to be employed for the physical activity and mental concentration are reconsidered in the following sections.

IV. CANDIDATE OBSERVABLE FEATURES FOR MENTAL CONCENTRATION

A. Distance from the Face to Work Objects

As described in section II.*C*, our preceding study employs the distance from the face of each trainee to the work surface as the observable feature for mental concentration. Since this observable feature does not contradict the result of morpheme analysis described in section III, the same observable feature is also considered for mental concentration in this article. However, this observable feature is not correctly measured in the preceding study but is approximated as the distance from the face to the front camera for simplicity. Thus, in this article, the distance from the face to work objects is properly measured without any approximation to use the distance as the observable feature of mental concentration.

However, to calculate this distance, two options can be considered as the position of the work objects. One is the position representing all the work objects on the work surface, and the other is that representing only the work objects being manipulated by hands. The former position can be obtained as $\hat{\boldsymbol{c}}(t)$, which has already been acquired in our preceding study, as described in II.*C*. The latter position can be obtained as the centroid of hand region $\mathcal{H}(t)$, because the manipulated objects are always with the hands. The position of this centroid is denoted by $\widehat{\boldsymbol{h}}(t)$. Here, if $\mathcal{H}(t)$ is obtained not as a single region but as two separated regions, $\mathcal{H}_1(t)$ and $\mathcal{H}_2(t)$, corresponding to different hands, the centroids of $\mathcal{H}_1(t)$ and $\mathcal{H}_2(t)$ are denoted by $\boldsymbol{h}_1(t)$ and $\boldsymbol{h}_2(t)$, and $\widehat{\boldsymbol{h}}(t)$ is defined as their average position.

Since $\hat{\boldsymbol{c}}(t)$ and $\widehat{\boldsymbol{h}}(t)$ are both represented by the 2D coordinate system associated with the image frames of the overhead camera whereas $\boldsymbol{f}^C(t)$ is represented by the front-camera-centered 3D coordinate system, they need to be transformed into the same 3D coordinate system to calculate the distances between them in 3D space. By calibrating the geometric relation between these two different coordinate systems via some reference points, such as corners on the work surface, in advance, $\hat{\boldsymbol{c}}(t)$, $\widehat{\boldsymbol{h}}(t)$, and $\boldsymbol{f}^C(t)$ can be transformed into 3D coordinates $\hat{\boldsymbol{c}}^W(t)$, $\widehat{\boldsymbol{h}}^W(t)$, and $\boldsymbol{f}^W(t)$ in the work-surface-centered coordinate system.

The distance from the face to work objects scattered on the work surface or manipulated objects among them can be evaluated by the Euclidean distance from $\boldsymbol{f}^W(t)$ to $\hat{\boldsymbol{c}}^W(t)$ or $\widehat{\boldsymbol{h}}^W(t)$, which is denoted by $\delta_c(t)$ $(= \|\hat{\boldsymbol{c}}^W(\tau) - \boldsymbol{f}^W(\tau)\|)$ or $\delta_h(t)$ $(= \|\widehat{\boldsymbol{h}}^W(\tau) - \boldsymbol{f}^W(\tau)\|)$. The average of this distance during any period $[t, t+\Delta t]$ in a video is considered a candidate observable feature for mental concentration in this article. This measurement denoted here by $\bar{\delta}_c[t, t+\Delta t]$ or $\overline{\delta_h}[t, t+\Delta t]$ is defined as follows:

$$\bar{\delta}_c[t, t+\Delta t] = \frac{1}{\Delta t}\sum_{\tau=t}^{t+\Delta t} \delta_c(\tau) \tag{1}$$

$$\bar{\delta}_h[t, t+\Delta t] = \frac{1}{\Delta t}\sum_{\tau=t}^{t+\Delta t} \delta_h(\tau) \tag{2}$$

B. Gaze at Work Objects

From the result in section III, it should also be considered a candidate observable feature whether each trainee is gazing at work objects or not. However, this is not considered in our preceding study. Thus, in this article, departure of the trainee's gaze from the direction towards work objects is considered one of the candidate observable features for mental concentration.

The unit normal vector of a trainee's face in the front-camera-centered coordinate system at *t* is denoted by $\boldsymbol{g}^C(t)$, which can be calculated from the face orientation

obtained in the process of estimating 3D face position $\boldsymbol{f}^C(t)$ in section II.*C*. Vector $\boldsymbol{g}^C(t)$ transformed into the work-surface-coordinate system is denoted by $\boldsymbol{g}^W(t)$. If the gaze direction can be assumed to coincide with the face normal, the departure of gaze from the direction towards working objects can be evaluated by the angle of $\boldsymbol{g}^W(t)$ from that direction. Similar to the discussion in section *A*, the destination of this direction can be specified by either $\hat{\boldsymbol{c}}^W(t)$ or $\hat{\boldsymbol{h}}^W(t)$ whereas the source of the direction is given only by $\boldsymbol{f}^W(t)$. For either of the two possible destinations, the angle of $\boldsymbol{g}^W(t)$ from $\hat{\boldsymbol{c}}^W(t) - \boldsymbol{f}^W(t)$ or $\hat{\boldsymbol{h}}^W(t) - \boldsymbol{f}^W(t)$, denoted here by $\theta_c(t)$ or $\theta_h(t)$ $(0 \le \theta_c(t),\ \theta_h(t) < \pi)$, can be considered. Thus, the average of this angle during any period $[t, t+\Delta t]$ in a video is introduced as another candidate observable feature for mental concentration. This measurement is denoted by $\bar{\theta}_c[t, t+\Delta t]$ or $\bar{\theta}_h[t, t+\Delta t]$, which is defined as follows:

$$\bar{\theta}_c[t, t+\Delta t] = \frac{1}{\Delta t}\sum_{\tau=t}^{t+\Delta t} \theta_c(\tau) \tag{3}$$

$$\bar{\theta}_h[t, t+\Delta t] = \frac{1}{\Delta t}\sum_{\tau=t}^{t+\Delta t} \theta_h(\tau) \tag{4}$$

Furthermore, this departure of gaze considered above and the distance from the face to work objects considered in section *A* can be combined into another candidate observable feature for mental concentration. However, the former is measured by an angle whereas the latter is measured by a distance. For a measurement reflecting both, the distance from $\hat{\boldsymbol{c}}^W(t)$ or $\hat{\boldsymbol{h}}^W(t)$ to the line passing through $\boldsymbol{f}^W(t)$ along with $\boldsymbol{g}^W(t)$ is employed. This distance is given simply as $\delta_c(t)\sin\theta_c(t)$ or $\delta_h(t)\sin\theta_h(t)$, which increases with $\delta_c(t)$ and $\theta_c(t)$, or $\delta_h(t)$ and $\theta_h(t)$ (see Fig. 4). Since this value begins to decreases after it attains $\delta_c(t)$ or $\delta_h(t)$ when $\theta_c(t)$ or $\theta_h(t)$ reaches $\pi/2$, $\delta_c(t)(2-\sin\theta_c(t))$ or $\delta_h(t)(2-\sin\theta_h(t))$ is measured instead for further increase of the value after that. In summary, this measurement, denoted here by $\Theta_c(t)$ or $\Theta_h(t)$, is defined as follows:

$$\Theta_c(t) = \begin{cases} \delta_c(t)\sin\theta_c(t) & (0 \le \theta_c < \pi/2) \\ \delta_c(t)(2-\sin\theta_c(t)) & (\pi/2 \le \theta_c < \pi) \end{cases} \tag{5}$$

$$\Theta_h(t) = \begin{cases} \delta_h(t)\sin\theta_h(t) & (0 \le \theta_h < \pi/2) \\ \delta_h(t)(2-\sin\theta_h(t)) & (\pi/2 \le \theta_h < \pi) \end{cases} \tag{6}$$

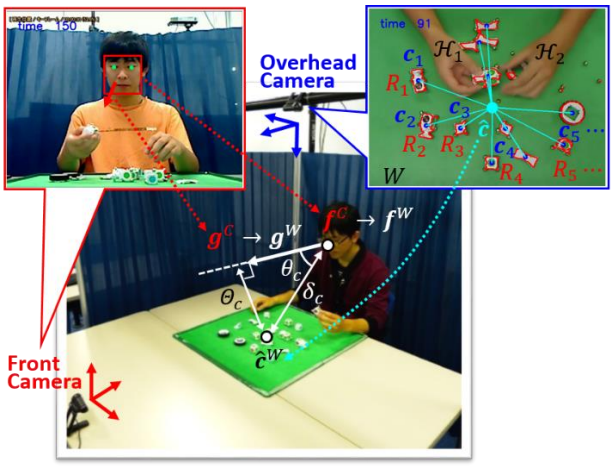


Fig. 4. Candidate observable features for the mental concentration for the position of work objects represented by $\hat{\boldsymbol{c}}^W(t)$.

Based on $\Theta_c(t)$ or $\Theta_h(t)$ defined above, its average during any period $[t, t+\Delta t]$ is considered another candidate observable feature for mental concentration. This measurement is denoted by $\bar{\Theta}_c[t, t+\Delta t]$ or $\bar{\Theta}_h[t, t+\Delta t]$, which is defined as follows:

$$\bar{\Theta}_c[t, t+\Delta t] = \frac{1}{\Delta t}\sum_{\tau=t}^{t+\Delta t} \Theta_c(\tau) \tag{7}$$

$$\bar{\Theta}_h[t, t+\Delta t] = \frac{1}{\Delta t}\sum_{\tau=t}^{t+\Delta t} \Theta_h(\tau) \tag{8}$$

C. Evaluation of Candidate Observable Features

To find the best observable feature that is most useful in estimating mental concentration categorized by humans from among possible candidates $\bar{\delta}_c$, $\bar{\delta}_h$, $\bar{\theta}_c$, $\bar{\theta}_h$, $\bar{\Theta}_c$, and $\bar{\Theta}_h$ considered in sections *A* and *B*, their correlations with the categories answered by the experimental participants for mental concentration in the video clips in step (2) in section III are evaluated. For facial image processing necessary for obtaining $\boldsymbol{f}^C(t)$ and $\boldsymbol{g}^C(t)$, *Face U* [17] supplied by *PUX Corporation* was employed. The resultant correlation coefficients are shown in Table II. In this result, $\bar{\delta}_c$, $\bar{\theta}_c$, and $\bar{\Theta}_c$, which employ $\hat{\boldsymbol{c}}^W$ as the position representing work objects, show better correlation than $\bar{\delta}_h$, $\bar{\theta}_h$, and $\bar{\Theta}_h$, employing $\hat{\boldsymbol{h}}^W(t)$ for the position. Since the largest absolute value of the correlation coefficient is given by $\bar{\Theta}_c$, it is found to be the best observable feature for the mental concentration in this article.

TABLE II: CORRELATION COEFFICIENTS OF CANDIDATE OBSERVABLE FEATURES FOR MENTAL CONCENTRATION

Position representing work objects	Candidate observable features	Correlation coefficients
$\hat{\boldsymbol{c}}^W$	$\bar{\delta}_c$	−0.73
	$\bar{\theta}_c$	−0.73
	$\bar{\Theta}_c$	**−0.76**
$\hat{\boldsymbol{h}}^W$	$\bar{\delta}_h$	−0.70
	$\bar{\theta}_c$	−0.72
	$\bar{\Theta}_h$	−0.75

V. CANDIDATE OBSERVABLE FEATURES FOR PHYSICAL ACTIVITY

A. Temporal Change or Decrease of Dispersion for the Positions of Work Objects

For the observable feature to estimate the physical activity, the amount or speed for operation of work objects seemed to be attended to as the observable feature by the experimental participants in the experimental result in section III. This observable feature can be evaluated as the temporal change in the dispersion for the positions of work objects. Similar to the discussion in section IV, the position of work objects can be represented by all the work objects on the work surface or only the objects being manipulated by hands. However, better correlation coefficients have already been obtained with the position represented by all the work objects in section III.*C*. Thus, only the position represented by all the work objects is considered in this section. Since the temporal change in the dispersion for the positions of all the work

objects has already been measured in our preceding work as $D[t, t+\Delta t]$, which more precisely denotes the average temporal change in the dispersion for the positions, this measurement is still considered as a candidate observable feature for the physical activity in this article.

Another observable feature that seemed to be attended to by the experimental participants in section III is work progress. Since work objects are in general assembled into a single object in assembly work, work progress can be characterized as the decrease of the dispersion in the position of work objects. Since the dispersion is obtained as $d(t)$ in our preceding study, as described in section II.*C*, the amount of its decrease is considered another candidate observable feature of the physical activity in this article. Total decrease of $d(t)$ for $[t, t+\Delta t]$, denoted by $D'[t, t+\Delta t]$, is defined as follows:

$$D'[t, t+\Delta t] = \sum_{\tau=t}^{t+\Delta t-1}\{d(\tau+1) - d(\tau)\} \quad (9)$$

B. Evaluation of Candidate Observable Features

Similar to section III.*C*, the usefulness of D and D' as the observable feature for the physical activity is evaluated based on their correlations with the categories answered by the experimental participants for the physical activity of the video clips in step (2) in section III. The resultant correlation coefficients are shown in Table III. In the result, D gives a larger value than D' for the correlation coefficient. Thus, D is introduced again as the best observable feature for the physical activity in this article.

TABLE III: CORRELATION COEFFICIENTS OF CANDIDATE OBSERVABLE FEATURES FOR THE PHYSICAL ACTIVITY

Candidate observable features	Correlation coefficients
D	**0.89**
D'	0.58

VI. ESTIMATING WORK SITUATION FROM THE BEST OBSERVABLE FEATURES

A. Precision for Estimating Work Situation

For estimating work situations with respect to the physical activity and mental concentration from their best observable features D and $\overline{\Theta}_c$ found in sections IV and V, the relation of those observable features with the categories of work situations answered by experimental participants in step (2) in section III is represented by a mathematical function. Here, the answer of each experimental participant for the category of work situation with respect to each conceptual attribute is represented by binary values 1 or 0, meaning *high* or *low*, for the attribute. The correct category of work situation for each video clip is determined as the average of the binary values for the answers of all the experimental participants. Those average values for the physical activity and mental concentration in the kth video clip ($k = 1, \cdots, 50$) are denoted by P_k and M_k, which take real numbers in $[0,1]$. The relations of D and $\overline{\Theta}_c$ with P_k and M_k are both approximated by a sigmoid function $s(\alpha, \beta; x)$ with parameters α and β defined as follows:

$$s(\alpha, \beta; x) = \frac{1}{1+e^{-\alpha(x+\beta)}} \quad (10)$$

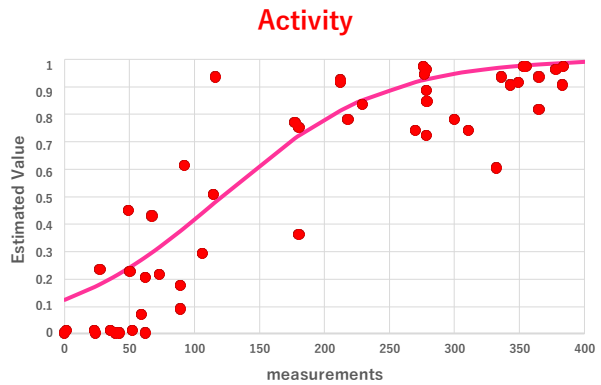
where $s(1,0; -\infty) = 0.0$, $(1,0; 0) = 0.5$, and $s(1,0; \infty) = 1.0$. Variable x is set as either D or $\overline{\Theta}_c$, depending on the conceptual attributes. Parameters α and β are determined so that the following error functions $E_P(\alpha, \beta)$ and $E_M(\alpha, \beta)$ are minimized:

$$E_P(\alpha, \beta) = \sum_{k=1}^{50}\|s(\alpha, \beta; D[t_k, t_k+\Delta t_k]) - P_k\|^2 \quad (11)$$

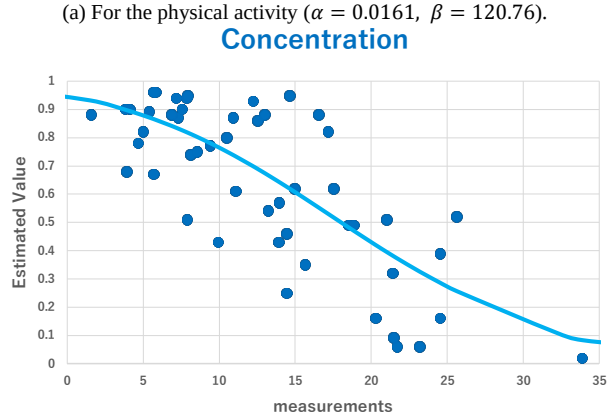
$$E_M(\alpha, \beta) = \sum_{k=1}^{50}\left\|s\left(\alpha, \beta; \overline{\delta}_c[t_k, t_k+\Delta t_k] + \overline{\Theta}_c[t_k, t_k+\Delta t_k]\right) - M_k\right\|^2 \quad (12)$$

where $[t_k, t_k+\Delta t_k]$ denotes the period corresponding to the kth video clip.

Fig. 5 illustrates sigmoid functions with the values for α and β obtained by minimizing $E_P(\alpha, \beta)$ and $E_M(\alpha, \beta)$ together with the values of P_k and M_k. Average errors for estimating P_k and M_k are 0.165 and 0.177. Since the categories of work situations answered by the experimental participants for each video clip are not the same but have fairly large variance (0.321 on average), these errors can be regarded as sufficiently small.



(a) For the physical activity ($\alpha = 0.0161$, $\beta = 120.76$).



(b) For the mental concentration ($\alpha = -0.1407$, $\beta = 17.472$).

Fig. 5. Approximation of the categories of work situations from the best observable features obtained from the video clips by sigmoid functions.

B. Possibilities of Producing Teaching Materials to Enhance Motivations

Finally, it is evaluated whether the video clips with the work situation estimated as *high activity & high concentration* actually enhance motivations of the viewers towards the assembly work more than those estimated as the

other three categories. Video clips estimated as each of the four categories are picked up from among all 50 video clips based on their values of the best observable features for the physical activity and mental concentration. The number of video clips picked up was 10 for *high activity & high concentration*, 7 for *high activity & low concentration*, 8 for *low activity & high concentration*, and 10 for *low activity & low concentration*. From among the videos picked up for each category, a single video clip of the front camera is randomly selected to form a quadruplet of video clips, each of which is estimated as a different category, as shown in Fig. 6.

Fig. 6. Sample image of a quadruplet of video clips presented to experimental participants for evaluating enhancement of their motivations towards the assembly work by the difference in work situation.

Ten experimental participants are presented with 10 different quadruplets of video clips to answer which of the four video clips in the presented quadruplet most enhances their motivations towards the assembly work. Table IV shows how many times the video clips of each category were chosen by the experimental participants in total. This result confirms that video clips with the work situations estimated as *high* for both the physical activity and mental concentration from their best observable features are actually effective for enhancing motivations of the viewers towards the assembly work.

TABLE IV: TOTAL NUMBER OF CHOICES FOR THE VIDEO CLIP ESTIMATED TO BE DIFFERENT CATEGORIES AS MOST MOTIVATIONAL

High activity & *High* concentration	**78**
High activity & *Low* concentration	16
Low activity & *High* concentration	6
Low activity & *Low* concentration	0

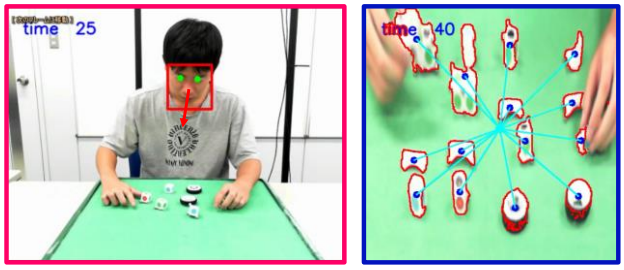
VII. CONCLUSIONS

This article discussed how to estimate work situations of practical training with assembly tasks with respect to the physical activity of the work and the mental concentration of each trainee on the work as the conceptual attributes for categorizing the situations from appropriate observable features obtainable by image processing for videos of the work. Our preceding study showed the possibility of estimating those work situations from some observable features, aiming to produce video previews that enhance motivations of viewers towards the assembly work by picking the scenes estimated as high physical activity and high mental concentration from videos taken in actual practical training classes. However, there remain several issues to discuss. First, it has not been discussed what our humans actually paid attention to as the observable features for discriminating work situations different with regard to the conceptual attributes. Second, precision for estimating work situations from appropriate observable features has not been quantitatively evaluated. Third, it has not been made clear whether videos with work situations estimated as high for both the physical activity and mental concentration actually enhance motivations of the viewers towards the assembly work.

This article has discussed the above three issues. For the first issue, the observable features that humans actually pay attention to for the two conceptual attributes are analyzed using morpheme analysis for the questionnaire among many experimental participants. Based on the result, candidate observable features are considered, and the best observable feature for each conceptual attribute is determined from the correlations of those candidate observable features with the correct categories answered by the experimental participants. The best observable features obtained as the result are the temporal change in the dispersion for the positions of all the work objects on the work surface as well as the sum of the distance from the face to the centroid of all the work objects and the departure of the gaze from the direction towards the centroid.

For the second issue, the average error between the work situations estimated from the best observable features and answered by the experimental participants is evaluated for each of the two conceptual attributes, and the result is much smaller than the variance of the answers by the experimental participants.

For the third issue, it was shown that video clips with work situations estimated as high for both the physical activity and mental concentration enhance motivations of the viewers much more than the others from the experiment where experimental participants are presented with video clips estimated as different work situations based on the best observable features to choose the most motivational ones.



(a) Camera image from a viewing direction upper than the front camera.

(b) Homography transformation of the work surface region in (a).

Fig. 6. Example of images of the face and the work surface obtained by a single camera by homography transformation.

Although the observable features proposed in this article are obtained from different cameras, it is also possible to use only a single camera to obtain those observable features by installing it so that it can observe both the face and the work surface. Fig. 6(b) is an example of the images obtained by transforming the image region of the work surface in a camera image in (a) by image processing called *homography transformation*. Images in (a) and (b) can be employed instead of those obtained by the front and overhead cameras

in this article. This extension of our method is considered as one of our possible future steps.

It should also be discussed as another future step whether the same approach can be applied for practical training of other kinds, because our work in this article only focuses on assembly work for practical training classes. Video teaching materials are most useful, especially for practical training, in which work progress can be easily recognized by simply observing the work from outside, and, thus, assembly work was focused on in this article as the most representative practical training where video teaching materials are effective. However, there should still be other kinds of practical work effective for introducing the approach discussed in this article.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 16H03225 and 26282062. We would like to thank Editage (www.editage.jp) for English language editing.

REFERENCES

[1] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, "Studying learning in the worldwide classroom: Research Into edX's first MOOC," *RPA Journal*, 2013.

[2] P. W. Dowrick, "Using video psychological and social applications," Chichester, UKWilly, pp.105–124, 1983.

[3] A. Bandura, "Toward a unifying theory of behavioral change," *Psychological Review*, vol.84, no.2, pp.191–215, 1977.

[4] D. H. Schunk, R. A. Hanson, and P. D. Cox, "Peer-model attributes and children's achievement behavior," *Journal of Educational Psychology*, vol. 79, pp. 54–61, 1987.

[5] D. H. Schunk and R. A. Hanson, "Self-modeling and children's cognitive skill learning," *Journal of Educational Psychology*, vol. 81, no. 2, pp. 155–163, 1989.

[6] C. H. Hitchcock, P. W. Dowrick, and M. A. Prater, "Video self-modeling intervention in school-based settings," *Remedial & Special Education*, vol. 24, p. 36, 2003.

[7] K. Okamoto, K. Kakusho, M. Yamamoto, T. Kojima, and M. Murakami, "Video-based performance recognition of assembly work in a practical training class for teaching material preparation," *Journal of Advances in Information Technology*, vol.7, no.3, pp. 186–193, 2016.

[8] K. Shutaro, M. Kazune, S. Hiroshi, and I. Hirotake, "The inference method of cognitive working states while performing mental tasks based on performance-cognitive load model," *International Conference on Human Computer Interaction*, vol. 17, no. 4, pp. 395–410, 2015.

[9] K. Uchiyama, K. Ooishi, K. Miyagi, H. Ishii, and H. Shimoda, "Process of evaluation index of intellectual productivity based on work concentration," presented at ICSTE 2013, 2013.

[10] C. Epp, M. Lippold, and R. Mandryk, "Identifying emotional states using keystroke dynamics," in *Proc. the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011)*, pp. 715–724, 2011.

[11] J. Fogarty *et al.*, "Predicting human interruptibility with sensors," *Computer-Human Interaction*, vol. 12, no. 1, pp. 119–146, 2005.

[12] R. Brunken, J. L. Plass, and D. Leutner, "Direct measurement of cognitive load in multimedia learning," *Educational Psychologist*, vol. 38, no. 1, pp. 53–61, 2003.

[13] P. Ayres and F. Paas, "Cognitive load theory: New directions and challenges," *Applied Cognitive Psychology*, vol. 26, no. 6, pp. 827–832, 2012.

[14] F. Paas, J. E. Tuovinen, H. Tabbersm, Pascal, and W. M. Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational Psychologist*, vol. 38, no. 1, pp. 63–71, 2003.

[15] F. Paas and J. J. G. Merri¨enboer, "The efficiency of instructional conditions: An approach to combine mental effort and performance measures," *Human Factors*, vol. 35, no. 4, pp. 737–743, 1993.

[16] MOSS robot. [Online]. Available: http://www.modrobotics.com/moss/

[17] U. Face. [Online]. Available: http://pux.co.jp/product/softsensor/faceu/

Kai Okaomoto was born in Japan on September 11, 1992. He received a B. Eng. in human system interaction from Kwansei Gakuin University in 2015. He is currently a postgraduate student in a master's program in the Graduate School of Science and Technology, Kwansei Gakuin University. His research interests include video image processing, especially for human behavior recognition.

Koh Kakusho received a B.Eng. degree in electrical engineering from Nagoya University and M.Eng. and Ph.D. degrees in communication engineering from Osaka University in 1988, 1990, and 1993, respectively. He is currently a professor in the Department of Human System Interaction at Kwansei Gakuin University. His major research interests include computer vision and image processing for recognizing or understanding various kinds of human behaviors observed in daily life environments.

Michiya Yamamoto received his B.E. in electrical engineering from Kyoto University in 1997 and M.E. and the doctoral degrees in energy science from Kyoto University in 1999 and 2002, respectively. Since 2002, he has worked as an assistant professor at Okayama Prefectural University. In 2009, he moved to Kwansei Gakuin University as an associate professor and became a professor at that university in 2015. His research interests are embodied interaction and communication. support.

Takatsugu Kojima is currently an associate professor in Faculty of Medicine, Shiga University of Medical Science, Japan. He received his Ph.D. in cognitive psychology from Kyoto University, Japan. His research interests include spatial cognition and language, non-verbal information in communication, and affective design.

Masayuki Murakami was born in Japan on July 4, 1973. He received a bachelor of integrated human studies from Kyoto University in 1997, the master of human and environmental studies from Kyoto University in 1999, and the PhD in informatics from Kyoto University in 2005. Currently, he is a professor in the Research Center for Multi-Media Education, Kyoto University of Foreign Studies. His research interests are educational technology and higher education. He is especially interested in how information communication technology can be used to analyze and improve teaching and learning.