Video-Based Performance Recognition of Assembly Work in a Practical Training Class for Teaching Material Preparation

Kai Okamoto, Koh Kakusho, and Michiya Yamamoto
School of Science and Technology, Kwansei Gakuin University, Sanda, Japan
Email: {kaisea, kakusho, michiya.yamamoto}@kwansei.ac.jp

Takatsugu Kojima
Faculty of Medicine, Shiga University of Medical Science, Otsu, Japan
Email: kojima@kojima-lab.net

Masayuki Murakami
Research Center for Multi-Media Education, Kyoto University of Foreign Studies, Kyoto, Japan
Email: masayuki@murakami-lab.org

Abstract—This article discusses how to recognize situations in which assembly work assigned in a class for a course in practical training is performed by each of the current participants with high activity and concentration based on video images of a class for preparation of video teaching materials to enhance the motivation of future course participants. In the discussion, observable features to be used as the visual clues for evaluating activity and concentration of the work by each participant in each scene of the video are proposed together with image processing for obtaining those features from the video images. Experimental results for videos of assembly work with different degrees of activity and concentration are also presented.

Index Terms—video teaching materials, assembly work, motivation enhancement, scene recognition, activity, concentration

I. INTRODUCTION

It is common in school education to use video teaching materials, especially through activities known as Massive Open Online Courses (MOOCs) [1]. Since it is troublesome to prepare video teaching materials by creating and editing videos manually, it has been proposed to prepare video teaching materials simply by photographing actual classes of various courses by installing video cameras in classrooms [2]-[5]. Although most such video teaching materials have been considered primarily for use in mass teaching through classroom lectures, in which the participants just sit while listening to lecturers talk, with the aim of acquiring various kinds of knowledge, it could also be useful to introduce video teaching materials for practical training, in which each participant is required to perform practical work independently to acquire a particular skill.

Although we can set various purposes for teaching materials, most conventional video teaching materials are prepared for the purpose of providing viewers with a clear understanding of new material and thus consist of visual scenes in which a lecturer talks about the topics of the lecture while presenting slides or writing on a blackboard. However, for a class on practical training, what must be explained to class participants are the procedures of the practical task each is required to perform. Those procedures can be explained fully by simply providing written directions, as in conventional classes for practical training. If video teaching materials could be made useful for practical training, in addition to written directions for the procedures of practical tasks, then emphasis could be placed on enhancing the motivation or interest of the participants towards the given practical task.

Not only for video teaching materials but also for learning processes in general, major factors affecting motivation to learn can be summarized in four concepts, attention, relevance, confidence, and satisfaction, which comprise the ARCS model in instructional design [6]. Since attention and relevance in our case are concerned with the attention of the learners to the learning contents and their relevance to learners’ interests, respectively, some of that attention and relevance can be assumed to be already present for class participants, because they have decided to join the class. On the other hand, satisfaction with the learning contents cannot be obtained simply by viewing video teaching materials regarding the practical work to be performed later, because the work to be done has not yet been conducted. As a result, if video teaching materials are to contribute to enhancing participants’ motivation in a class for practical training, the
contribution must be concerned with the remaining factor, inspiring confidence in accomplishing the required practical work.

Another previous study points out that learner confidence arises from self-efficacy [7], which originates in performance experience, verbal persuasion, vicarious experience, imaginal experience, and psychological and emotional states. From among these sources, the most possible result of viewing video teaching materials of practical work is vicarious experience of the work. As a result of the discussions above, video teaching materials for enhancing the motivation of participants in a class for practical training must include scenes in which the required practical work is performed well by a participant other than the viewers of the video teaching materials.

In order for the viewers of video teaching materials to recognize that the practical work is performed well by a participant in the video, the video needs to cause visible change during its performance. The practical work that most satisfies this requirement and thus is suitable for introducing video teaching materials is assembly work, because it is performed through changes in the layout of the assembled parts. In fact, many educational television programs for children deal with various kinds of assembly work as their topics.

In this article, we discuss how to recognize a scene in which a participant in a class for practical training performs the required practical work well from a video obtained by photographing a class, with the aim of preparing video teaching materials for enhancing the motivations of viewers as future participants in the same course by editing scenes of the video. The remainder of this article is organized as follows. In section 2, we discuss the conceptual attributes for categorizing various practical work situations with respect to performance of the work, as well as observable features useful for recognizing the situations in each category from video images of the work. In section 3, we propose an image processing procedure for obtaining those observable features for assembly work from video images. Some experimental results for evaluating the possibility of recognizing situations of assembly work categorized by the conceptual attributes based on the observable features obtained by image processing for sample videos of actual assembly work are presented in section 4. Concluding remarks and discussion of our future work are provided in section 5.

II. CONCEPTUAL Attributes FOR CATEGORIZATION AND OBSERVABLE FEATURES FOR RECOGNITION

A. Conceptual Attributes for Categorizing Human Work Performance

In research on the methodology for evaluating human intellectual work productivity, productivity is evaluated primarily with respect to two different aspects of the work, physical activity and the worker's mental concentration. Many previous studies have proposed focusing on physical activity to evaluate work productivity using personal computers (PCs) by analyzing each worker’s operation of a PC [8]-[10]. On the other hand, previous studies considering workers’ cognitive loads have tried to estimate productivity based on workers’ concentrations or tensions while performing work, by using cognitive load theory [11] [12]. There is also previous study that used both of these physical and mental features to evaluate the intellectual productivity of a worker [13] [14].

Referring to the previous investigations mentioned above, we also used physical activity during work and mental concentration of each participant in the class for practical training as the conceptual attributes to categorize the performance of the practical work. In this paper, physical progress is defined as how actively each participant performs the required practical work, whereas mental concentration means how deeply the participant is devoted to the work with high concentration. We used these two attributes as the axes to categorize practical work situations for each participant with respect to his or her performance. Quadrants I, II, III, and IV, as defined by these two axes, correspond to the following situations.

1) (High Activity & High Concentration)

The required practical work is performed actively by the participant concentrating highly on the work; this situation is ideal for education.

2) (High Activity & Low Concentration)

The required practical work is performed actively, but the participant is not concentrating on the work. This situation often occurs when a participant is forced to perform work despite not interested in it.

3) (Low Activity & Low Concentration)

The required practical work is not performed actively by the participant, and the participant does not concentrate on the work. This situation often occurs when a participant is neither interested in the work nor forced to work.

4) (Low Activity & High Concentration)

The required practical work is not performed actively even though the participant is concentrating on the work. This situation often occurs when a participant is not familiar with the work.

Fig. 1 shows sample images of typical scenes corresponding to these situations. In the remainder of this section, we will discuss observable features for evaluating the situations with respect to the conceptual features corresponding to these two axes.

Figure 1. Sample images of situations corresponding to quadrants I–IV.
B. Observable Feature to Evaluate Activity of Assembly Work

As described in section 1, previous studies on evaluating activity in work focused primarily on work performed on PCs. Since each work operation can be obtained and recorded easily for work on a PC, if necessary, the work activity can be evaluated based on the work amount or the operation type as the observable feature.

If we could obtain each operation also for assembly work, we could evaluate the work activity based on the work amount or the operation type, similarly to work on a PC. However, unlike work on a PC, each operation cannot be obtained easily for assembly work through observation by a video camera. It is necessary to track each part of the work surface completely throughout the assembly process in order to obtain each operation, but tracking often fails as a result of occlusion due to connection and disconnection of the parts to be tracked and manipulation of the parts by human hands.

Instead of obtaining each operation for the parts, we focus on the dispersion of the positions of the parts scattered around the work surface. Since any assembly work involves assembling various parts into a single object consisting of those parts, the number of objects existing on the work surface must change during the assembly process. Moreover, the parts are all separated and scattered widely around the work surface at the beginning of the assembly work, and then some of them are gathered into one place when they are combined to form a larger block. As a result, the dispersion of the positions of the objects on the work surface actively changes while the assembly work is being performed.

Fig. 2 shows sample images of the process of assembling a robot from its parts. The dispersion changes frequently when the activity during the assembly work is high, but it remains similar when the activity is low. Thus, we used the total change in the dispersion of the positions of parts on the work surface as the observable feature to evaluate activity in assembly work.

High Activity

Low Activity

Figure 2. Sample images of workplace with the progress of assembly work.

C. Observable Feature to Evaluate Worker’s Concentration on Assembly Work

Previous work has shown that physiological indices such as heart rates and pupil diameters, which can be obtained by heartbeat meters and eyeball sensors, respectively, are useful for evaluating the degree of human concentration on work [15]-[17]. However, it is not realistic to attach those sensors to the bodies of each of the participants in a class for practical training, because that could impede the participants’ learning process. Moreover, what needs to be recognized in our work is not the degree to which each participant is actually concentrating on his or her practical work, but the degree of concentration estimated by the viewers of the video of the participants. The participants’ concentrations as estimated by the viewers must be recognizable only from the features observable in the video.

Previous studies on evaluating human concentration on specific tasks on PCs have proposed using observable features obtainable without attaching specific sensors to workers. The most representative features include the distance between the face of the worker and the display of his or her PC [18]-[20]. This observable feature could also be useful for work in the physical world when PCs are not involved. Fig. 3 shows sample images of a person performing assembly work. When the person is leaning forward during the work, the distance between her face and the object being assembled decreases, and the image gives the impression that this person is performing the work with high concentration, whereas the image of the person leaning backward with a long distance between her face and the work object during her performance of the work gives the impression of low concentration.

High Concentration

Low Concentration

Figure 3. Sample images of person with different distances from work objects.

Based on this fact, we used the distance between the face of the participant and the work objects as the observable feature to evaluate concentration on the work. More precisely, we actually used the distance of the face from the work surface as the observable feature for simplicity, because work objects are usually placed on work surfaces unless their sizes are sufficiently small.

In the next section, we will describe how to obtain the observable features proposed above from video images through image processing.

III. Obtaining Observable Features From Video Images

A. Measuring Total Change in Dispersion of Positions of Work Objects on Work Surface

We evaluated the dispersions of the position of the work objects on the work surface by extracting the image regions containing those work objects from each image frame. For this image processing discussion, we assume that the camera is installed so that it looked down on the
work surface from straight above. We also assume that the color of the work surface is different from the colors used for the work objects, enabling the participant to distinguish each of them easily from the work surface. In this setting, image features for evaluating the dispersion of the positions of the work objects can be obtained from each image frame as follows (see Fig. 4).

The reason why the variance of the region centroids is not used, but rather the sum of their squared deviations, is to consider the effect of the number of work objects separately scattered on the work surface. Since the variance is the result of dividing the sum of the squared deviations by the number of object regions, that value is not affected by their number. This value can remain large until the very last stage of assembly, when all the parts have already been assembled into two large blocks just before completion of assembly, as far as the distance between those two blocks from each other is long enough. After those two blocks are assembled into a single object by the last operation of assembly, the value rapidly decreases to zero. The sum of their squared deviations is used instead of the variance of the region centroids to evaluate the dispersion of the object positions to avoid this inconvenient property.

The total change in the dispersion during a particular period $[t, t + Δt]$ in the process of assembly is denoted by $D[t, t + Δt]$. It is given as the sum of the absolute temporal difference of $d(t)$ at each moment $t$ over all the moments in that period, defined as follows:

$$D[t, t + Δt] = \sum_{t=1}^{t+Δt-1} |d(τ + 1) − d(τ)|$$

B. Measuring Distance between Face and Camera Installed on Work Surface

To measure the distance between the face of the participant and the work surface using a video camera, it is necessary to use the camera to obtain video images of the participant’s face. In this discussion of the method used to obtain those video images, it is assumed that another camera is installed on the work surface to observe the participant from the front, in addition to the camera installed for observing the work surface, as described in the previous section. The image frame obtained by this camera at moment $t$ is denoted by $f(t)$. In this camera setting, the distance of the face to the work surface can be approximated by the distance to this camera. Thus, the distance between the face of the participant and the camera is used instead of that between the face and the work surface.

It becomes easy to extract facial regions together with various facial features corresponding to facial parts, including the eyes and mouth, as well as to estimate the 3D pose of each extracted face based on the appearance of its facial region together with facial parts, using recent facial image processing techniques. As a result of such facial image processing for $f(t)$, the 2D position of the person’s eyes in the image and the 3D pose of his or her face relative to the camera can be obtained at each moment $t$ of the video. Fig. 5 illustrates a sample facial image processing result.

$$\hat{c}(t) = \frac{1}{K} \sum_{k=1}^{K} c_k(t)$$

where $c_k(t)$ is the average position of $c_1(t), \cdots, c_{K}(t)$, defined as follows:

$$d(t) = \sum_{k=1}^{K} |c_k(t) − \hat{c}(t)||^2$$

where $\hat{c}(t)$ is the average position of $c_1(t), \cdots, c_{K}(t)$.
In this discussion, a camera-centered coordinate system is assumed, with its origin at the optical center of the camera, the z-axis along the optical axis from the back to the front of the camera, and the x- and y-axes parallel to the image plane from the left to the right and from the bottom to the top of the camera, respectively. If the 3D position of the face can be obtained in this camera-centered coordinate system, the distance of the face to the camera is the z coordinate of the 3D position. This 3D position can be estimated from the 2D positions of the eyes and the 3D orientation of the face obtained as the result of the facial image processing described above (see Fig. 6). The 2D homogeneous coordinates of the 2D positions of the left and right eyes on the image plane of the camera-centered coordinate system for image frame $f(t)$ are defined by $l^f(t)$ and $r^f(t)$, respectively.

**Figure 5.** Facial region and positions of eyes extracted from image captured by camera installed in front of person performing assembly work on work surface.

**Figure 6.** Geometric relations of positions of face and camera in camera-centered and face-centered coordinate systems.

In addition to the camera-centered coordinate system described above, we also discuss a face-centered coordinate system with the origin at the midpoint between the eyes, the x-axis passing through the eyes from the left to the right of the face, the z-axis directed perpendicular to the surface of the face from back to front, and the y-axis in the direction upward from the face. The 3D position of the face in the camera-centered coordinate system corresponds to the 3D position of the origin of the face-centered coordinate system. This 3D position at moment $t$ is denoted by $f^c(t)$. The 3D pose of the face in the camera-centered coordinate system determines the rotation for the transformation of the face-centered coordinate system to the camera-centered coordinate system. The matrix representing this rotation at moment $t$ is denoted by $\Phi(t)$. If the coordinates of the 3D positions of the left and right eyes at moment $t$ in the face-centered coordinate system are denoted by $l^f(t)$ and $r^f(t)$, respectively, and those in the camera-centered coordinate system by $l^c(t)$ and $r^c(t)$, respectively, their geometric relations can be represented by $f^c(t)$ and $\Phi(t)$, respectively, as follows:

$$\begin{align*}
\{ l^f(t) &= \Phi(t)l^c(t) + f^c(t) \\
r^f(t) &= \Phi(t)r^c(t) + f^c(t)
\end{align*}$$

Since $l^f(t)$ and $r^f(t)$, which represent the 2D positions of the left eye and the right eye on the image plane, are the result of optical projection of $l^c(t)$ and $r^c(t)$ onto the image plane in the camera-centered coordinate system, they can be represented as follows:

$$\begin{align*}
\{ l^f(t) &= \lambda P[\Phi(t)l^c(t) + f^c(t)] \\
r^f(t) &= \lambda P[\Phi(t)r^c(t) + f^c(t)]
\end{align*}$$

where $P$ is the matrix representing the process of optical projection of the 3D space onto the image plane, and $\lambda$ is the scaling parameter.

In Eq. (8) can be obtained by using strong camera calibration in advance. The positions of $l^c(t)$ and $r^c(t)$ can be set as $l^c(t) = (-\delta/2, 0, 0)^T$ and $r^c(t) = (\delta/2, 0, 0)^T$, respectively, where $\delta$ denotes the standard human interocular distance, because human interocular distances do not vary substantially by individual. As a result, $f^c(t)$, which represents the position of the face at moment $t$, is the only unknown variable in Eq. (8). Thus, $f^c(t)$, can be obtained by solving these equations.

The $z$ coordinate of $f^c(t)$, denoted here by $f^c_z(t)$, serves as the distance from the face to the camera. The concentration of the participants performing the activity and the concentration on the work, respectively, would correctly reflect the impression of the viewers of videos of assembly work of the activity and the concentration of the participants performing the assembly work in the videos, through experiments using video images obtained by photographing actual assembly work. We used the MOSS robot [21] produced by

Modular Robotics for the assembly work. This product consists of parts that mostly have the same cubic shape with a magnet at each corner, and thus various kinds of robots can be built from the same set of parts simply by physically connecting them with each other in different formations without cabling (see Fig. 7).

We asked six participants to build robots that could move by themselves using at least one sensor, which was also provided as one of the parts. To obtain video images of situations in different categories for the assembly work, we divided the six participants into two groups of three participants each, and assigned them different tasks. We asked each participant in the first group, first, to build a robot after providing instructions on how to build the robot, expecting to create situations in categories IV and I. We allowed each participant in each group an intermission of 150 s after each task, expecting to create situations in category III.

We prepared the work surface as a plastic green mat in the shape of 650 mm × 650 mm square with a short bank at each edge to prevent the parts to be assembled from rolling out of the mat during assembly. We installed two cameras in front of and directly over the work surface at distances of 120 cm and 90 cm, respectively, from the work surface and the participant, in order to observe the work surface and the participant, as shown in Fig. 8. Fig. 1-Fig. 3 were actually obtained using those cameras.

We segmented the video images obtained under the settings described above into 10-s-long video clips and selected situations that provide impressions corresponding to categories I-IV. For each of the video clips selected and categorized above, we calculated the image features \( D[t, t + \Delta t] \) and \( Z[t, t + \Delta t] \) during the period \([t, t + \Delta t] \) corresponding to the video clip by setting \( \Delta t \) equal to 10 s. The values of \( H_{\min}^W, H_{\min}^H, H_{\max}^W, H_{\max}^H, s_{\min}^W, s_{\min}^H, s_{\max}^W, s_{\max}^H, \theta \) and \( \delta \) were set as \( H_{\min}^W = 80, H_{\max}^W = 160, H_{\max}^H = 80, S_{\min}^W = 40, S_{\max}^W = 255, S_{\max}^H = 255, \theta = 300 \) and \( \delta = 63 \). We employed Face U supplied by PUX Corporation for facial image processing [22].

**Experimental Results**

We segmented the video images obtained under the settings described above into 10-s-long video clips and selected situations that provide impressions corresponding to categories I-IV. For each of the video clips selected and categorized above, we calculated the image features \( D[t, t + \Delta t] \) and \( Z[t, t + \Delta t] \) during the period \([t, t + \Delta t] \) corresponding to the video clip by setting \( \Delta t \) equal to 10 s. The values of \( H_{\min}^W, H_{\min}^H, H_{\max}^W, H_{\max}^H, s_{\min}^W, s_{\min}^H, s_{\max}^W, s_{\max}^H, \theta \) and \( \delta \) were set as \( H_{\min}^W = 80, H_{\max}^W = 160, H_{\max}^H = 80, S_{\min}^W = 40, S_{\max}^W = 255, S_{\max}^H = 255, \theta = 300 \) and \( \delta = 63 \). We employed Face U supplied by PUX Corporation for facial image processing [22].

**Fig. 9 illustrates the two measurements plotted on the 2D surface with the horizontal and vertical axes representing \( Z \) and \( D \), respectively. The values obtained for the video clips classified into different categories are shown with different types of symbols. These results show that the obtained measurements, which represent the total dispersion of the positions of the working objects and the distance of the face to the camera on the work surface, reflect the differences among categories I-IV with respect to the conceptual attributes, namely, activity and concentration on the work. From these results, it is evident that these measurements well reflect these conceptual attributes.**

**V. CONCLUSIONS**

In this article, we discussed how to recognize assembly work situations with high participant activity and high concentration on the work, aiming to produce video teaching materials that enhance the motivation of future
participants in a practical training course involving assembly work by selecting those situations from a video obtained by photographing an actual class of the course. Referring to knowledge from previous studies on factors enhancing learner motivation regarding learning contents, we focused on the performance of the work by the current class participants appearing in the video teaching materials. We introduced two conceptual attributes, activity and concentration on the work, to classify situations with different degrees of performance based on previous work on evaluating the productivity of intellectual work. To estimate the degrees of the two conceptual attributes from an assembly work video, we proposed using the total dispersions of the positions of the work objects as the observable features. Moreover, we also proposed how to measure those observable features through image processing and showed that those measurements accurately reflected the conceptual attributes in an experiment using videos of actual assembly work performed by some participants. Since this experiment used video images of assembly work performed by only six participants, and since the categorizations of the video clips were determined from the viewpoint of the authors, it is necessary to use different numbers of participants for the assembly work and to have the video clips categorized by other participants as one future step of this investigation. It is also essential to develop a process of recognition and selection of assembly work situations with high activity and concentration based on the measurements proposed in this article.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 26282062. The authors would like to thank Enago (www.enago.jp) for the English language review.

REFERENCES


Kai Okaomoto was born in Japan, on September 11, 1992. He received B. Eng. in Human System Interaction from Kwansei Gakuin University in 2015. He is currently a Postgraduate student of 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 the B.Eng. degree in electrical engineering from Nagoya University, 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 BE in electrical engineering from Kyoto University in 1997, ME and doctoral degrees in energy science from Kyoto University in 1999 and 2002, respectively. Since 2002, he has been working 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, and so on.

Takatsugu Kojima is currently an associate professor in the Faculty of Medicine, Shiga University of Medical Science, Japan. He received his PhD in Cognitive Psychology from Kyoto University, Japan. His research interest includes spatial cognition and language, non-verbal information in communication, and affective design.

Masayuki Murakami was born in Japan, on July 4, 1973. He received Bachelor of Integrated Human Studies from Kyoto University in 1997, Master of Human and Environmental Studies from Kyoto University in 1999 and 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. Especially, he is interested in how they can use information communication technology to analyze and improve teaching and learning.