The Classification of Different Situations in a Lecture Based on Students' Observed Postures

Yuki Kotakehara[1(✉)], Koh Kakusho[1], Satoshi Nishiguchi[2], Masaaki Iiyama[3], and Masayuki Murakami[4]

[1] Kwansei Gakuin University, 2-1 Gakuen, Sanda 669-1337, Japan
{yuki-kotakehara, kakusho}@kwansei.ac.jp
[2] Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata 573-0196, Japan
satoshi.nishiguchi@oit.ac.jp
[3] Kyoto University, Yoshidahonmachi, Sakyo-ku, Kyoto 606-8501, Japan
iiyama@mm.media.kyoto-u.ac.jp
[4] Kyoto University of Foreign Studies, 6 Saiinkasamecho, Ukyo-ku, Kyoto 615-8558, Japan
masayuki@murakami-lab.org

Abstract. This paper discusses the possibility of identifying different situations related to the students during a lecture from its video by classifying the situations that happen in the lecture based on the similarity in the posture of each student. The recognized situations can be used as indexes for the instructor to watch the video to further improve the lecture. Although it has been shown in a previous work that there are some relations between the postures taken by the students and their understanding of the lecture, it is not clear what types of situations actually happen during the lectures, and the postures taken by the students differ even when they are in the same situations. To deal with these problems, the representative postures of each student in different situations are first obtained by clustering the postures actually taken by the student, and then different situations of the class are obtained by clustering the combinations of representative postures of all the students under the assumptions that similar postures are taken by each student and similar combinations of those postures are observed for the whole group of students when they are in the same situation.

Keywords: Lecture situation · Student posture · Clustering

1 Introduction

Recently, in the field of higher education, it has been often suggested to record the lectures on videos to review them for *Faculty Development* (FD) [1–3]. However, the instructors cannot easily select the scenes to be watched, and it is very time-consuming for them to review their lectures by watching the whole video. To reduce the heavy workload of watching the lecture videos, previous works have proposed to recognize various situations related to the instructor and the students for indexing the videos [4–9].

P.-L. P. Rau (Ed.): HCII 2019, LNCS 11577, pp. 446–457, 2019.
https://doi.org/10.1007/978-3-030-22580-3_33

Those previous works can be classified into two types: those that consider mainly the situations related to the instructor [4–7], and those that focus on the students [8, 9]. The previous works of the first type discuss how to recognize the instructor's behaviors, which include writing on the blackboard, presenting slides, talking to the students and so on. Those of the second type focus mainly on students' behaviors because it has been pointed out that there is a relation between students' behaviors and their interest during the lectures. That is, students' behavior of looking ahead often reflects their interest in the lecture [10].

Additionally, recent work has analyzed the relation between the postures taken by the students during a lecture, and as the result, it has been shown that different behaviors such as dozing off and looking away as well as looking ahead can be used as useful clues to estimate the students' understanding of the lecture [11]. Based on these results, this article discusses how to recognize combinations of those behaviors of the whole group of the students in the classroom during a lecture as the situation of the lecture. To this aim, it is necessary to clarify what kinds of situations can be observed in the lecture, because the situations of lectures related to the behaviors of the whole group of the students are not so well organized as those of the instructors, who gives the lectures with the specific purpose of giving clear explanations using slides and whiteboards. Moreover, whereas most students look ahead when they are paying attention to the lectures, the postures taken by the students while they are dozing off or looking away might be different for different students.

In our work, we classify different types of situations from the combinations of the behaviors observed for the whole group of the students at different moments of the lectures. To cope with individual differences in the postures for the same behavior in this classification, we assume that the same posture taken by the same student implies the same behavior of the student, and classify different behaviors of each student based on the similarity between the postures actually observed for the student. More precisely, first we obtain representative postures for each student by clustering his/her postures observed at each moment of the lecture. Then, we describe specific situations at each moment of the lecture combining the representative postures of all students attending the lecture. Finally, those situations are again clustered based on the similarity in the combination of the representative postures, and different situations related to the students during the lecture are recognized.

In Sect. 2, we will provide a more detailed explanation of the procedure used in this study. In Sect. 3, we will present the results of an experiment conducted by one of the authors in his university to evaluate the procedure described above. Finally, in Sect. 4, we will summarize the main points of this article and discuss possible future steps for our research.

2 The Classification of Students' Situations by Clustering Their Postures

2.1 The Identification of Representative Postures for Each Student

The posture of each student observed in each frame of the lecture video can be obtained by conventional human image processing techniques for pose estimation. The obtained posture is described by the two-dimensional (2D) coordinates of all the observable feature points of the student's body. Let $\boldsymbol{x}_i(t)$ denote the posture of i-th student denoted by S_i observed in t-th frame denoted by F_t of the lecture video ($i = 1, \cdots, N$; $t = 1, \cdots, T$), where N and T denote the number of the students observed in the lecture video and that of the frames constituting the video, respectively. The posture $\boldsymbol{x}_i(t)$ is a $2J$ dimensional vector, where J denotes the number of feature points, mainly the joints, of a student's body. In this article, this vector is named the *observed posture* of student S_i at frame F_t. Since each observed posture describes only 2D positions in the image frame for the feature points of each student, and therefore does not include any information concerning depth, the observed posture changes according to the geometric relation between the student and the camera used to take the lecture video, even when the same posture and the same student are involved. However, it is possible to keep this geometric relation unchanged by fixing the camera in the classroom, given that each student sits in the same seat throughout the lecture. Under this condition, the difference in observed posture $\boldsymbol{x}_i(t)$ reflects the difference in actual 3D posture of student S_i.

The set of all the observed postures in each video frame obtained for student S_i is denoted by $O_i = \{\boldsymbol{x}_i(1), \cdots, \boldsymbol{x}_i(T)\}$. Assuming that each student should take similar postures for the same behavior, the clusters denoted by $C_i = \left\{C_i^1, \cdots, C_i^{K(i)}\right\}$, in which $K(i)$ denotes the number of the clusters, are obtained by grouping all the postures included in O_i, which should correspond to the number of different postures actually taken by student S_i, and thus differs from one student to another (see Fig. 1). Since the observed postures $\left\{\boldsymbol{x}_i(t) | \boldsymbol{x}_i(t) \in C_i^{k(i)}, C_i^{k(i)} \in C_i\right\}$, which are all classified into the same cluster $C_i^{k(i)}$, are similar to each other, those postures are regarded as representing the same posture taken by student S_i for the same behavior. The representative postures are defined to indicate these observed postures taken for the same behavior by each student. The $k(i)$-th *representative posture* $\boldsymbol{X}_i^{k(i)}$ of student S_i is defined by the centroid of $C_i^{k(i)}$ as follows:

$$\boldsymbol{X}_i^{k(i)} = \frac{1}{|C_i^{k(i)}|} \sum_{\boldsymbol{x}_i(t) \in C_i^{k(i)}} \boldsymbol{x}_i(t) \tag{1}$$

where all the representative postures of student S_i are given by the set $X_i = \left\{\boldsymbol{X}_i^1, \cdots, \boldsymbol{X}_i^{K(i)}\right\}$.

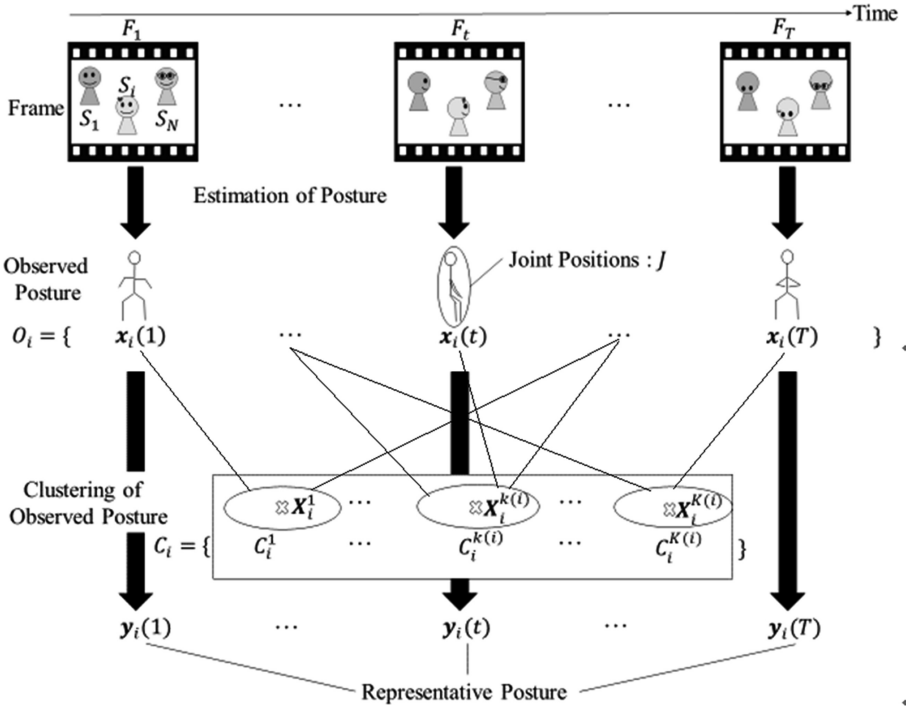


Fig. 1. Representative postures obtained by clustering observed postures.

To describe the behavior associated to an observed posture for each student at any frame, the observed posture is substituted by the representative posture that is more similar to that observed among all the representative postures of the student. Let $\boldsymbol{y}_i(t)$ denote the representative posture to substitute observed posture $\boldsymbol{x}_i(t)$ of student S_i in frame F_t. This representative posture is given as that with the minimal Euclidian distance from $\boldsymbol{x}_i(t)$ in X_i as follows:

$$\boldsymbol{y}_i(t) = \underset{\boldsymbol{X}_i^{k(i)} \in X_i}{\operatorname{argmin}} \left\| \boldsymbol{x}_i(t) - \boldsymbol{X}_i^{k(i)} \right\| \tag{2}$$

2.2 The Classification of Different Situations in the Whole Group of Students

As a result of the procedure described in Sect 2.1, representative postures $\boldsymbol{y}_1(t), \cdots, \boldsymbol{y}_N(t)$ of all the students $S_1, \cdots, S_N$ are obtained for each frame F_t. Since any observed posture is described as a $2J$ dimensional vector, any of the N representative postures are also described as a $2J$ dimensional vector. These N representative postures are employed to describe the situation of the whole group of students in each frame. The situation of the whole group of students in frame F_t is denoted by $\boldsymbol{y}(t)$, which is

called here *combined representative posture*, and it is defined as the 2*JN* dimensional vector, whose elements are constituted by those of the *N* representative postures as follows:

$$\boldsymbol{y}(t) = [\boldsymbol{y}_1(t) \cdots \boldsymbol{y}_N(t)] \tag{3}$$

Let *R* denote the set of the combined representative posture $\boldsymbol{y}(t)$ for all the frames, where $R = \{\boldsymbol{y}(1), \cdots, \boldsymbol{y}(T)\}$. Since the frames in which each student takes the observable postures to be substituted by the same representative posture of his/her own should be regarded as the frames with the same behavior for the whole group of the students, the frames with similar combined representative postures can be regarded as the frames representing the same situation for the whole group of students. Based on this idea, the sets of all combined representative postures *R* are classified into the clusters, each including similar combined representative postures (see Fig. 2). The resultant set of clusters is denoted by $D = \{D^1, \cdots, D^L\}$, where *L* is the number of clusters corresponding to the number of different situations that actually occurred during the observed lecture.

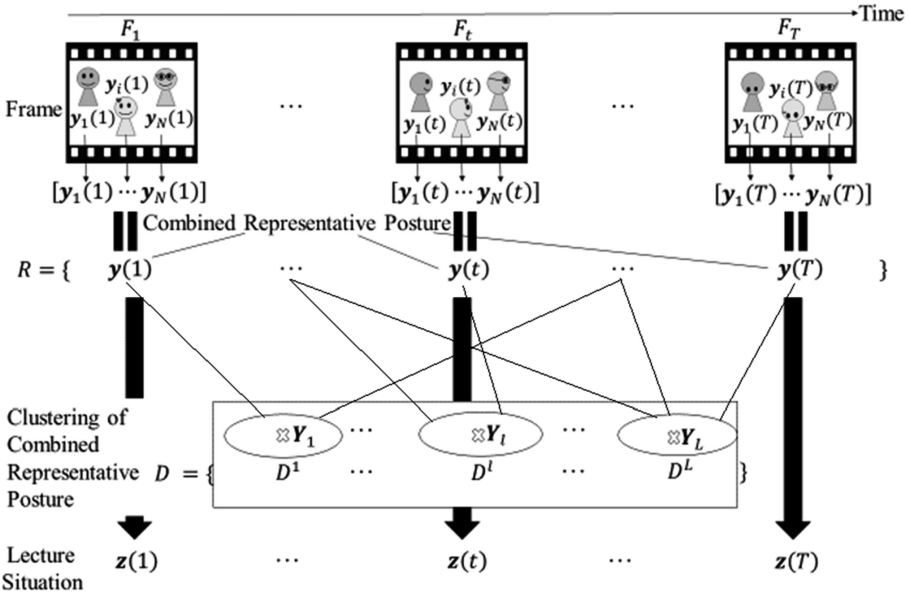


Fig. 2. Lecture situations obtained by clustering representative postures.

If the situation for each frame needs to be further recognized among its possible variations obtained as *D* described above, the situation to be recognized for frame F_t

can be obtained by replacing combined representative posture $\boldsymbol{y}(t)$ with the centroid of the cluster including $\boldsymbol{y}(t)$. Let $\boldsymbol{Y}_l$ denote the centroid of cluster D^l $(l = 1, \cdots, L)$, which is defined as follows:

$$\boldsymbol{Y}_l = \frac{1}{|D^l|} \sum_{\boldsymbol{y}(t) \in D^l} \boldsymbol{y}(t) \tag{4}$$

where $Y = \{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_L\}$ describes different situations of the whole group of the students. Thus, the situation of the whole group of students in frame F_t can be recognized by finding $z(t)$, which denotes the element with the minimal Euclidean distance from $\boldsymbol{y}(t)$ among Y:

$$z(t) = \underset{\boldsymbol{Y}_l \in Y}{\operatorname{argmin}} \|\boldsymbol{y}(t) - \boldsymbol{Y}_l\| \tag{5}$$

Since Y is not given in advance but is obtained based on the similarity between the students' postures, in order to identify the situations in which the students are involved, we do not need to know in advance neither what kinds of situations possibly happen during the lecture nor what postures are actually taken by each student in each situation.

3 Experimental Results

3.1 Students' Observed Postures

We run an experiment to evaluate whether the method described in Sect. 2 can be successfully used to identify situations that are useful for instructors to review and improve their lectures. We recorded the seminar supervised by one of the authors of this article by fixing a camera in the classroom after obtaining students' approval. The recorded video consisted of 2771 frames (T = 2771) and lasted 90 min. The results of pose estimation for the students appearing in the video included the postures of 13 students out of all those who attended the seminar for each frame (N = 13). OpenPose [12] was employed to pose estimations. Postures of all the other students could not be obtained due to occlusions among the students. The observed posture for a student for each frame is described as a 24-dimensional vector, which consists of 2D coordinates in the image frame for 12 feature points, including the nose, neck, shoulders, elbows, wrists, eyes, and ears (J = 12). Figure 3 illustrates the observed postures for the 13 students in a frame of the lecture video. Different lines indicate different pairs of feature points adjacent to each other. The face of each student is hidden in the image for privacy protection.

Fig. 3. An example of the observed postures.

3.2 Representative Postures Obtained for Each Student

The observed postures obtained for each student in all frames were classified into clusters of similar postures. The *k-means* method [13] was employed for clustering. Since this method requires that the number of clusters $K(i)$ is specified, we tried different values for $K(i)$ in order to find the appropriate number for the clusters. As a result, clusters including the observed postures that can be interpreted as meaningful behaviors were obtained for $K(i) = 2–8$.

Figures 4 and 5 show examples of the observed postures included in each of the three clusters obtained for two different students when $K(i) = 3$. The observed postures in this example can be interpreted as the behaviors of *looking ahead*, *taking notes*, and *looking away*. However, the observed postures included in the clusters corresponding to the same behavior for different students are not necessarily similar in terms of their geometric shapes. This result implies that the observed postures taken by different students during the same lecture may have a similar variation of their behavior, whereas the geometric shapes of the observed postures that can be interpreted as the same behavior often include individual difference. Nevertheless, our method allows us to extract meaningful behaviors that occur during the lecture while tolerating individual differences in the observed postures by merely clustering the observed postures of each student.

(a) Examples of observed postures for looking ahead.

(b) Examples of observed postures for taking notes.

(c) Examples of observed postures for looking away.

Fig. 4. The representative postures of student A.

3.3 Obtaining the Situations of the Whole Group of Students

The representative postures of each student were obtained as the centroids of the clusters of the observed postures obtained in Sect 3.2 to replace the observed postures of the student in each frame with one of those representative postures and form the combined representative postures for the whole group of students in the frame. By clustering the combined representative postures in all frames, different situations of the group of students during the lecture were obtained. The *k-means* method was employed again for clustering. Since the number of clusters L is unknown, we tried different values also for L. As a result, most clusters could be interpreted as meaningful situations for the whole group of students for $L = 4$.

Figures 6 and 7 show examples of frames classified into different clusters. In each figure, the representative posture of each student is shown at the position of the student in the image frame. Figure 6 shows examples of situations that can be given a meaningful interpretation, whereas the situations depicted in Fig. 7 cannot be interpreted meaningfully. For example, the situations illustrated in Fig. 6 can be interpreted respectively as (a) *paying attention to the lecture*, (b) *taking notes*, and (c) *looking*

(a) Examples of observed postures for looking ahead.

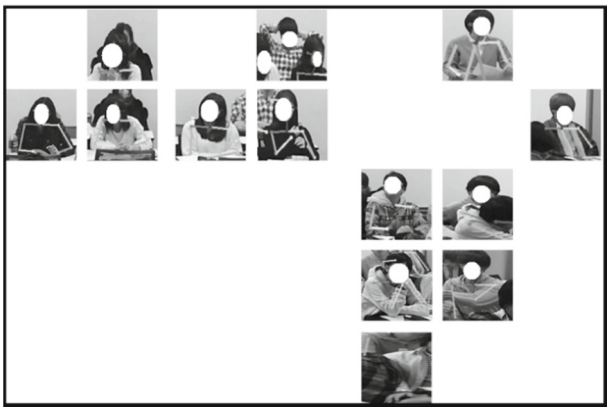
(b) Example of observed postures for taking notes.

(c) Examples of observed postures for looking away.

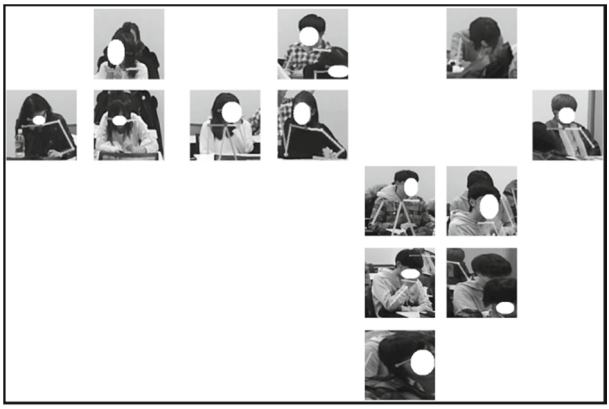
Fig. 5. The representative postures of student B.

away, because almost all students show the same behavior although the geometric shapes of the representative postures are different. On the other hand, the examples in Fig. 7 are not easily interpreted in a meaningful way for the whole group, because some students are paying attention to the lecture while others are taking notes.

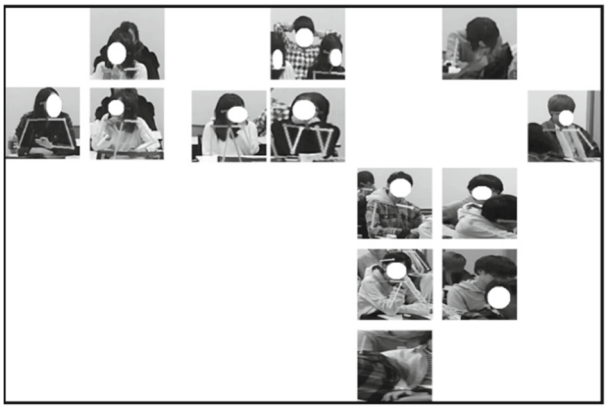
From the examples reported above, it can be said that our method is fairly useful in obtaining meaningful situations of the students regardless of their individual differences in posture, but still needs further improvement. One of the reasons why the situations in Fig. 7 cannot be interpreted univocally is that the students begin and finish taking notes in different moments. To deal with the asynchrony of the behaviors, it is necessary to make our clustering method tolerant for a slight temporal difference.



(a) Sample situations to be interpreted as *looking ahead*.



(b) Sample situations to be interpreted as *taking notes*.



(c) Sample situations to be interpreted as *looking away*.

Fig. 6. Examples of situations that have meaningful interpretations.

Fig. 7. Examples of situations that are difficult to interpret meaningfully.

4 Conclusions

This article discussed the possibility of identifying various situations related to the whole group of students during lectures from the videos obtained with a fixed camera in the classroom. The proposed method first obtains observed postures for each student, described as 2D positions of the feature points of the body, by pose estimation for each frame of the recorded lecture. Since each student is seated at the same location throughout the lecture and the camera is fixed in the classroom, the differences in the observed postures of each student reflect the changes in his/her posture. Thus, assuming that the same posture of the same student reflects the same behavior, the observed postures of each student in all the frames are classified into clusters based on their similarity to obtain the representative postures as the centroids of the clusters. The representative postures of all students in each frame are used to form the combined representative postures in the frame, and different situations of the whole group of students during the lecture are obtained by further clustering the combined representative postures in all the frames. Applying this method to the analysis of the video of a seminar, most of the obtained clusters could be given meaningful interpretations, although some of them were difficult to interpret meaningfully.

In future research, we need to modify the method so that the clustering can tolerate individual differences related to the moment in which the posture changes. Although the most straightforward solution would be to reduce the temporal resolution of the video frames, further discussion is required to understand how to address this issue properly.

It is also important to consider the different relevance of different feature points for evaluating the similarity between different postures based on their positions. For example, the position of each hand is not as relevant as the position of the head for evaluating the similarity in the posture of the whole body, because the hands tend to take more different positions than the head for the same behavior including paying attention, taking notes, and looking away. Thus, it becomes necessary for the clustering

to give different weights to different feature points or to normalize the distance between the feature points for evaluating the difference in posture.

References

1. Minoh, M., Nishiguchi, S.: Environmental media – in the case of lecture archiving system. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS (LNAI), vol. 2774, pp. 1070–1076. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45226-3_146
2. Coursera. https://www.coursera.org. Accessed 10 Dec 2018
3. edX. https://www.edx.org. Accessed 10 Dec 2018
4. Onishi, M., Fukunaga, K.: Shooting the lecture scene using computer-controlled cameras based on situation understanding and evaluation of video images. In: International Conference on Pattern Recognition (ICPR), Cambridge, pp. 781–784. IEEE (2004)
5. Shimada, A., Suganuma, A., Taniguchi, R.: Automatic camera control system for a distant lecture based on estimation of teacher's behavior. In: IASTED International Conference on Computers and Advanced Technology in Education, pp. 106–111 (2004)
6. Yousaf, M.H., Azhar, K., Sial, H.A.: A novel vision based approach for instructor's performance and behavior analysis. In: International Conference on Communications, Signal Processing, and their Applications (ICCSPA), Sharjah, pp. 1–6. IEEE (2015)
7. Lin, Y.-T., Tsai, H.-Y., Chang, C.-H., Lee, G.C.: Learning-focused structuring for blackboard lecture videos. In: Fourth International Conference on Semantic Computing, Pittsburgh, pp. 149–155. IEEE (2010)
8. Narayanan, S.A., Prasanth, M., Mohan, P., Kaimal, M.R., Bijlani, K.: Attention analysis in e-learning environment using a simple web camera. In: International Conference on Technology Enhanced Education (ICTEE), Kerala, pp. 1–4. IEEE (2012)
9. Yongyi, C.: Construction of a course video resource system based on students' visual attention. In: International Conference on E-Business and E-Government, Guangzhou, pp. 3840–3844. IEEE (2010)
10. Murakami, M., Kakusho, K., Minoh, M.: Analysis of students' eye movement in relation to contents of multimedia lecture. In: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn), pp. 1965–1968 (2002)
11. Mukunoki, M., Yoshitsugu, K., Minoh, M.: Students' posture sequence estimation using spatio-temporal constraints. In: Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R.R. (eds.) IPMU 2012. CCIS, vol. 298, pp. 415–424. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31715-6_44
12. Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, pp. 1302–1310. IEEE (2017)
13. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)