Intelligent Methods in Engineering Sciences 2(3): 084-089, 2023

IMIENS

INTELLIGENT METHODS
IN ENGINEERING SCIENCES

https://www.imiens.org

International
Open Access

September, 2023

e-ISSN 2979-9236

Research Article

https://doi.org/10.58190/imiens.2023.47

Visualization of Interest Level using OpenPose with Class Videos

Kang Dongshik [a, *], ***Yoshiaki Sasazawa*** [b], ***Minoru Kobayashi*** [c]

[a] *University Of The Ryukyus Faculty Of Engineering, Japan*
[b] *University Of The Ryukyus Faculty Of Education, Japan*
[c] *Bunkyo University Faculty Of Education, Japan*

ARTICLE INFO

Article history:
Received 27 July 2023
Accepted 28 September 2023

Keywords:
interest level,
visualization,
OpenPose,
classroom video,
face direction

ABSTRACT

The approach from the teacher's side to improve the classroom is to establish a learning discipline and an improvement method. However, there are various ways to do this, and it is an important part of teachers' work. In this paper, we propose a system for estimating the level of interest from class videos as a means of knowing the learning status of individual students. Using OpenPose, the system detects a person in a class video and extracts feature data of his/her joints. And we measure the concentration level based on the information of posture, facial orientation, and other movements. In addition, assuming that a person's face is facing the direction of the target when he/she is concentrating, we measure the concentration level based on the information of posture, facial direction, and other movements.

1. INTRODUCTION

Since the beginning of the Japanese school education system, the demonstration lesson has mainly been subjectively evaluated by the subject, and there has been almost no objective evaluation of attitudes in class. New evaluation methods are being explored. The approach from the teacher's side to improve the classroom is to establish a learning discipline and a supportive culture. Examples of how to speak and how to listen can be given as examples of instruction to ensure thorough study discipline. For example, students should listen facing the direction of the person speaking, and they should speak facing the direction of the person listening. If the interest level of each student can be estimated from the information on their posture, facial direction, and other behaviors, teachers will be able to use the results to structure their lessons to make it easier for students to concentrate.

In a previous research by Nishimura et al.[1], sensors were installed to students, and posture estimation was performed based on the recorded sensor information. However, there is a lack of objectivity and special sensors may interfere with the concentration of the students. With a camera image, it is possible to obtain necessary information without disturbing the concentration of students. In addition, it is easy to introduce the system to the educational field from the viewpoint of cost. In a previous study by Muramatsu et al.[2], the interest level of students in remote classes was measured from facial images and typing speed, but in face-to-face classes, students mainly use notebooks and handouts instead of PCs, so it is difficult to measure the interest level from typing speed. In previous research in our laboratory, we used OpenCV to detect students from classroom videos taken by a fixed camera installed in the classroom and determine their face direction. However, the OpenCV-based person detector often incorrectly detects other objects than persons as persons, and it is difficult to accurately determine the face direction.

In this study, as a means to know the learning state of each student, we estimate the interest level from videos taken by a camera installed in a classroom during the class. By introducing Openpose[3,4], it is possible to detect people and extract joint feature data, and calculate the interest level from posture and face direction information. The increase or decrease values of the coordinates of both eyes between frames is calculated, and the occurrence of motion is calculated by comparing them with a threshold value. From these results, a time series of motion records for all students are created, and the interest level is estimated by the proposed formula.

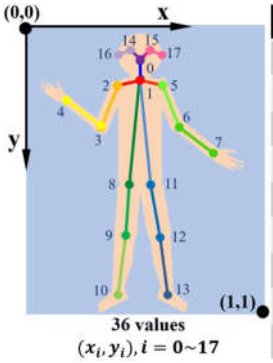
2. OpenPose

OpenPose[3,4] is a posture estimation library that extracts keypoints of joints of multiple persons from monocular camera images in real time using deep learning. The keypoints are 18 joints and include information not

* **Corresponding Author:** kang@ie.u-ryukyu.ac.jp

only from the bodies of the persons in the video, but also from their faces, hands, and legs. Figure 1 shows the detected keypoints.

OpenPose converts input images into features in advance, and calculates a Confidence Map to estimate key points of human joints and Part Affinity Fields to show the relationship between key points. The posture is then estimated by matching the keypoints obtained from the Confidence Map with the keypoints obtained from the Part Affinity Fields. Each keypoint is represented by a 2D vector, and the x-coordinate and y-coordinate of the keypoints that cannot be detected are set to 0.



No.	Key point	No.	Key point
0	Nose	9	Right knee
1	Neck	10	Right ankle
2	Right shoulder	11	Left hip
3	Right elbow	12	Left knee
4	Right wrist	13	Left ankle
5	Left shoulder	14	Right eye
6	Left elbow	15	Left eye
7	Left wrist	16	Right ear
8	Right hip	17	Left ear

Figure 1 Key points detected using OpenPose[5]

3. Proposal Systems

In this chapter, the extraction of people and their interest level using OpenPose are visualized, based on the video taken by a camera installed in a classroom. The proposed system consists of five major parts: video input part, person (student) detection part, motion detection part, interest level estimation part, and output part. The processing process of the proposed system is shown in Figure 2.

3.1. Video input part

The class format is for a classroom lecture in which all students are seated. The input (classroom) video is taken by a single fixed-point camera in the front of the entire classroom. In order to reduce the processing time in OpenPose, one frame per second of the classroom video is cut out, and then inputted.

3.2. Human detection part

For input video data, OpenPose calculates 18 joint coordinates. From among them, a human region is generated by adding the width and height (Δx and Δy) in an arbitrary range with the neck coordinates as the center. Then, the region for the number of persons is generated, and the coordinates of the necks that exist within the region are determined to be the center of the person. The detected human area is assigned to the person from the front row of the classroom. The values of Δx and Δy for determining the human area are determined by measuring the number of pixels in the width and height of the person's face, and are set to be approximately 2.5 to 3 times the value of Δx and Δy.

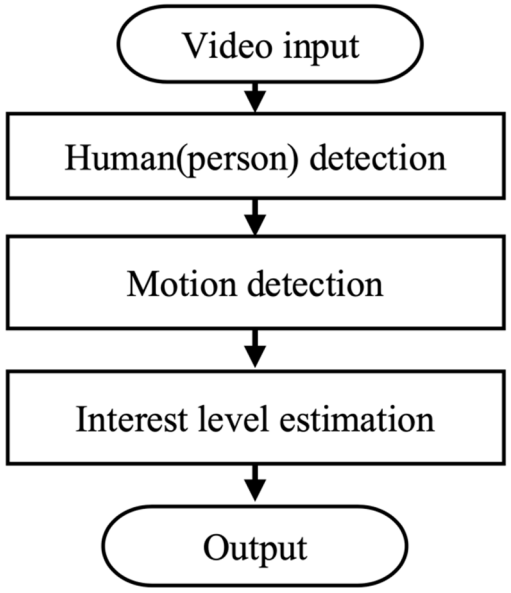


Figure 2 Flow Chart

3.3. Motion detection part

In this paper, we perform three types of motion detection. They are "facial motion," "facial direction," and "upper body motion" with the coordinates of both eyes, both shoulders, and the neck.

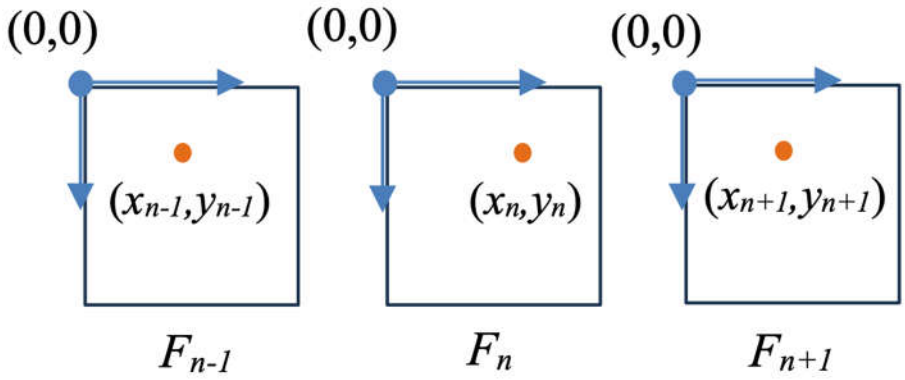


Figure 3 Facial motion between two frames

3.3.1. Facial motion

Facial motion is detected using changes in the *x*- or *y*-coordinates of the eyes. As shown in Figure 3, two frames are taken, one before and one after the other, and the coordinates of the eyes in the previous frame are x_{n-1} and y_{n-1}, and the coordinates of the eyes in the current frame are represented by x_n and y_n. The difference values of *x* and *y* are obtained by Equation (1), respectively. When this difference value exceeds a certain threshold value, it is judged that a motion has occurred. The threshold value is based on the pixel width of each individual's face.

$$D_x(n) = F_{n-1}(x_{n-1}) - F_n(x_n)$$
$$D_y(n) = F_{n-1}(y_{n-1}) - F_n(y_n) \quad (1)$$

3.3.2. Facial direction

To estimate the change in face direction, the coordinates of both eyes and the neck are used. As shown in Figure 4, given the coordinates A for the right eye, B for the neck, and C for the left eye, let ∠ABC be θ. The measure of θ is given by Equation (2).

$$\theta = \cos^{-1}\left(\frac{\overrightarrow{BA}\cdot\overrightarrow{BC}}{|\overrightarrow{BA}|\cdot|\overrightarrow{BC}|}\right) \quad (0 < \theta < 180) \quad (2)$$

When the orientation of the face changes, the measure of θ changes. The value of θ is maximized when the face is facing the camera. However, since the frontal position that the participant should see varies depending on the position of the seat in relation to the camera and the position of the lecturer, we assume that the average value of θ obtained from all frames is the orientation of the face looking at the frontal position.

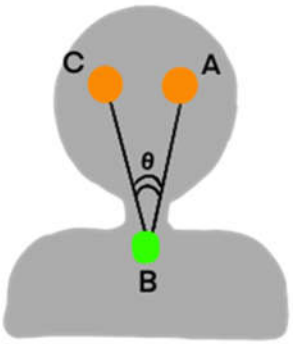


Figure 4 Facial direction

3.3.3. Upper body motion

Facial motion is detected although the face is facing the front, it is possible that the upper body is facing in a different direction from the front. Therefore, we estimated the orientation of the upper body, assuming that the subjects are more concentrated when their upper body and face are facing the same direction. The coordinates of both shoulders and neck are used to estimate the change of upper body motion. As shown in Figure 5, when the coordinates of the right shoulder A, neck B, and left shoulder C are given, let θ denote ∠ABC. The measure of θ is given by Equation (2). The upper body is assumed to be the frontal plane with θ within the range of the mean and mode values obtained from all frames for each student.

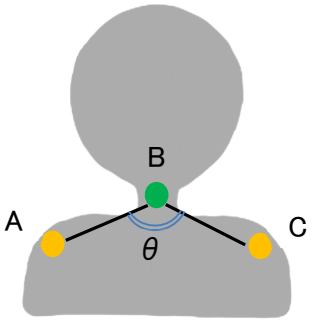


Figure 5 Upper body motion

3.4. Interest level estimation part

Facial motion is detected although the face is facing the front, it is possible that the upper body is facing in a different direction. In this part, the overall concentration of the student at time *t* is $C_{level}(t)$, the concentration obtained from the transition of facial movements is $C_{face}(t)$, the concentration obtained from the left and right directions of the face is $C_{direction}(t)$, and the concentration obtained from the transition of upper body movements is $C_{body}(t)$. The overall concentration $C_{level}(t)$ is calculated based on Equation (3). When the subject is facing the target direction, 1 is added to $C_{face}(t)$ and $C_{body}(t)$, and *-1* is added to $C_{face}(t)$ and $C_{body}(t)$ when there is motion from the front.

If the target direction is facing, *1* is added to $C_{direction}(t)$, and if the target direction is not facing, *-1* is added to $C_{direction}(t)$.

The state in which the value of $C_{level}(t)$ is high is estimated as the state in which the concentration level is high. In addition, the environment assumed here is a classroom lecture, and all the participants are seated and do not move around. Therefore, it is expected that the students themselves change their posture in some way, such as changing their posture or looking away from the classroom, when an action occurs.

$$C_{level}(t) = C_{level}(t-1) + C_{face}(t) + C_{direction}(t) + C_{dody}(t) \quad (3)$$

4. Experiments and Results

In this chapter, we describe the experiments verified on the basis of the proposed system.

4.1. Experiment environment

In this experiment, we used the video of a special class about sleeping and education. The classroom was brightly

lighted by fluorescent lamps and sunlight from the windows, and there were no displays that resembled human bodies, such as portraits or posters. A GoPRO camera was used, and it was placed at one location above the whiteboard in the front of the classroom. The camera was set at a resolution of 1,440p at fps 60.

The class consisted of a 50-minute lecture on sleeping using slides, and was conducted in a classroom style where the participants sat and listened to the lecturer. In the first 30 minutes, the lecturer explained about sleep, and in the latter 20 minutes, the participants filled out a form provided by the lecturer. The input video used in this experiment is a clip of the first 10 minutes after the lecture begins. The GPU of Google Colaboratory[6] was used to run OpenPose, and Python was used to analyze the data.

4.2. Human detection

The first frame of the outputted video from OpenPose is shown in Figure 6. Thirty-one persons were detected out of a total of 32 persons including 30 persons in the classroom and the person walking in the hallway, which is a detection rate of 96.8 %. However, two of the 31 detected persons are outsiders who are not the students. The non-subjects are eliminated from the sample. The two persons for whom missing values were so large that we could not obtain reliable coordinate values are excluded from the experiment. After excluding them, the 27 participants are the subjects of this experiment. Since all the participants in this study wore masks, there were many occasions when missing time periods were observed due to the influence of the masks.

In the case of the unmasked subjects, the influence of masks does not exist, but there were some subjects showed many false detections due to the temporary overlap with the previous subject.

Figure 6 Classroom video

4.3. Motion detection

In this chapter, we describe three types of motion detection form classroom video.

4.3.1. Facial motion detection

Figure 7 shows an example result of the x-coordinate graphs of both eyes and neck. The transition of facial motion in the frame is detected from the distance of movement of x and y coordinates of both eyes and neck in the previous and next frames using Equation (1). In this experiment, there were some frames in which keypoints could not be detected because the participants were wearing masks. Since the undetectable frames do not affect the results of the experiment, we treat them as missing values.

In addition, the difference of the movement distance between the right eye and the left eye was not detected in OpenPose because some students turned to either the left or right direction and turned their profile toward the camera. Or the two persons overlapped each other or the mask caused false detections in OpenPose.

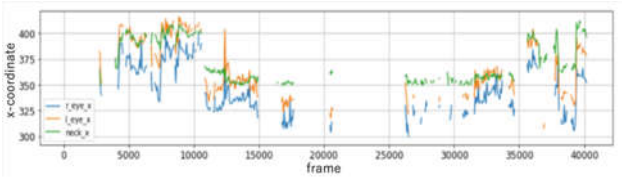


Figure 7 Result of facial motion detection in human ID-1

4.3.2. Face direction detection

The transitions of the face and upper body movements were obtained using Equation (2), but the frontal orientation of the participant changes depending on the location of the monitor. Therefore, we estimated the left and right directions in addition to the directions of the face and upper body. The left and right orientations were estimated by the difference between the coordinates of the two eyes and the threshold value. The average of the x-coordinates of each participant's eyes is used as the threshold value.

Figure 8 shows the time series of the orientation of human ID-1. The values of *1* and *-1* indicate the left and right orientations, respectively. 0 indicate missing values.

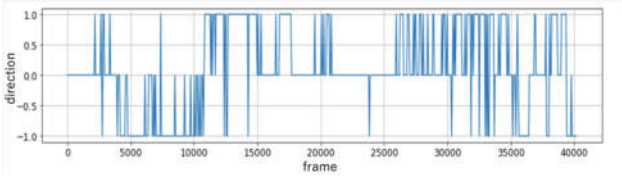


Figure 8 Result of facial motion detection in human ID-1

4.3.3. Upper body motion detection

Figure 9 shows a graph of x-coordinates of both shoulders and neck. Using Equation (2), we detected the transition of the movement of the upper body in the frame based on the distance of the x- and y-coordinates of the shoulders and neck in the front and rear frames. In this experiment, there were some frames in which key points could not be detected because the students wore masks. Since the undetectable frames do not affect the results of the experiment, we treat them as missing values. In addition, undetected frames and false positives similar to those detected in according with the subsection *4.3.1* were observed.

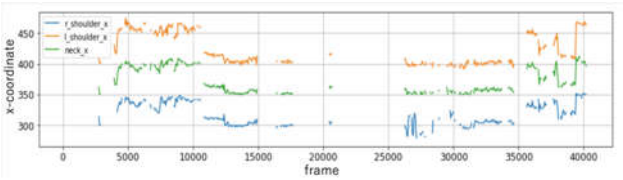


Figure 9 Result of upper body motion detection in human ID-1

4.3.4. Interest level estimation

The interest level for each individual participant is estimated based on Equation (3). Figure 10 shows the interest level estimated based on facial movements, Figure 11 shows the interest level after correction for the upper body motion, and Figure 12 shows the interest level estimated from the left and right direction of face. The red dotted line represents the average of the concentration for each direction.

Comparison of Figure 10 and 11 shows that the orientation of the face and the upper body tend to be similar. And Figure 13 shows the concentration of human ID-1 by time series.

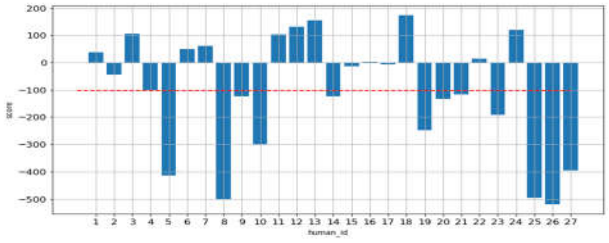


Figure 10 Interest level was estimated based on facial movements in human ID-1

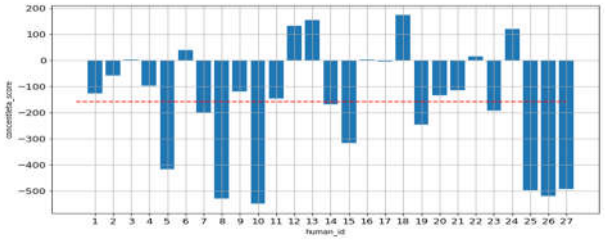


Figure 11 Interest level was estimated based on upper body motion in human ID-1

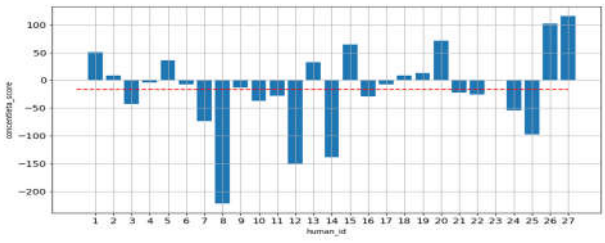


Figure 12 Interest level was estimated based on facial direction in human ID-1

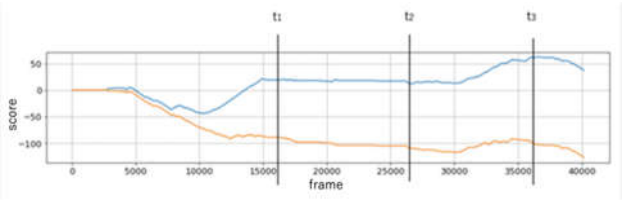


Figure 13 Concentration scores in human ID-1

4.3.5. Comparison Interest level and comprehension

In order to confirm whether there is a correlation between the concentration level calculated from the postural information and the comprehension level, a questionnaire in the form of a test was administered in this experiment. In this test, the participants were asked to choose one answer from a list of 12 options. Figure 14 shows the heat maps of the concentration and comprehension estimated from the concentration of only the face orientation, the concentration after correction for the upper body orientation, and the left and right orientations. f_score is the concentration of face orientation only, b_score is the concentration after correction for body orientation, and v_score is the concentration estimated from the left and right direction of face.

Figure 14 shows that the correlation with comprehension level is higher in the order of left-right orientation, upper body orientation, and face orientation. In addition, it is found that there is a high correlation between the estimation method with only the face orientation and the estimation method with the upper body orientation.

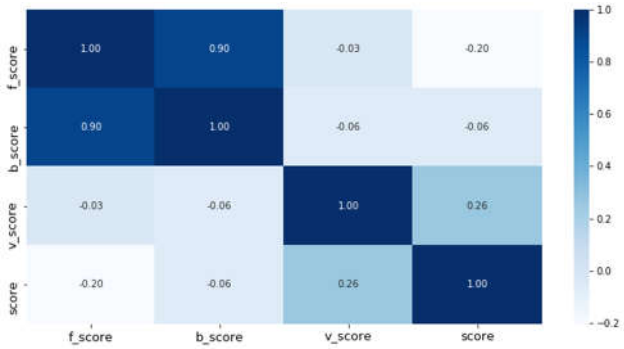


Figure 14 Heat maps of the concentration and comprehension

5. Conclusions

In this experiment, all the participants wore masks, so there were many frames in which key points could not be detected. In addition, it is essential to minimize the overlap between participants when using video for concentration estimation. In the concentration estimation section, we proposed a system for estimating the concentration level based on actions alone, but we believe that more accurate estimation of the concentration level will be possible by measuring the number of quizzes after the class and the number of active actions during the class.

In addition, although we estimated the concentration level only by the actions of the students in this experiment, we believe that it is necessary to take the actions of the instructor into consideration as well. Since frames for which keypoints could not be detected were treated as missing values, there was room for improvement in accuracy.

Acknowledgments

We would like to express our sincere gratitude to the teachers and students of junior high schools in Okinawa Prefecture for cooperation in this research. And this work is supported by the JSPS KAKEN Foundation JP22K02892.

Author's Note

This paper was presented at 11th International Conference on Advanced Technologies (ICAT'23), 17-19 August 2023, Istanbul, Turkey.

References

[1] Yukimasa NISHIMURA, Yoshito TOBE,“Suggestion of simple wireless sensors measure the concentration of teaching,” *The 16th National Convention of IEICE*, 2011.

[2] Tatsuma Muramatsu, Akihiko Sugiura,“Effect Verification of Concentration Measurement System that Uses Face Informations,” *the 74th National Convention of IPSJ*, 2012.

[3] Z.Cao. Hidalgo, T. Simon, S.Wei, “Y.Sheikh:Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” arXiv preprint ar Xiv:1812.0808

[4] Zho Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields, *In CVPR*, 2017.

[5] Yunkai Zhang, Yinghong Tian, Pingyi Wu, Dongfan Chen, “Application of Skeleton Data and Long Short-Term Memory in Action Recognition of Children with Autism Spectrum Disorder,” *Sensors 2021*, 21, 411.

[6] Google, “Google Colaboratory,” *https://colab.google*, 28 September 2023.