Estimating Positions of Students in a Classroom from Camera Images Captured by the Lecturer's PC

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

[1] School of Science and Technology, Kwansei Gakuin University, Sanda, Japan
{jun-nishikawa, kakusho}@kwansei.ac.jp
[2] Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan
iiyama@mm.media.kyoto-u.ac.jp
[3] Faculty of Information Science and Technology, Osaka Institute of Technology, Hirakata, Japan
satoshi.nishiguchi@oit.ac.jp
[4] Research Center for Multimedia Education, Kyoto University of Foreign Studies, Kyoto, Japan
masayuki@murakami-lab.org

Abstract. We propose to estimate the position of each student in a classroom by observing the classroom with a camera attached on the notebook or tablet PC of the lecturer. The position of each student in the classroom is useful to keep observing his/her learning behavior as well as taking attendance, continuously during the lecture. Although there are many previous works on estimating positions of humans from camera images in the field of computer vision, the arrangement of humans in a classroom is quite different from usual scenes. Since students in a classroom sit on closely-spaced seats, they appear with many overlaps among their regions in camera images. To cope with this difficulty, we keep observing students to capture their faces once they appear, and recover the positions in the classroom with the geometric constraint that requires those positions to be distributed on the same plane parallel to the floor.

Keywords: Classroom · Position estimation · Student positions · Continuous observation

1 Introduction

In the field of higher education, it becomes quite usual to record videos of lectures [1–3]. Those videos are used for various purposes including self-learning by students, reviewing the recorded lectures for faculty development, and so on. In addition to simply recording videos of lectures, it is also proposed to recognize various situations of the lecturer or the students during each lecture by observing it with cameras or sensors installed in the classroom. Those recognized situations are assumed to be

N. Streitz and P. Markopoulos (Eds.): DAPI 2015, LNCS 9189, pp. 518–526, 2015.
DOI: 10.1007/978-3-319-20804-6_47

employed as indices to find remarkable scenes from the recorded video for viewing the video or analyzing the procedures of the lecture efficiently [4–7].

In this article, we propose to estimate the position of each student sitting in a classroom during a lecture by observing the classroom with a camera, as one of the situations of the lecture. Positions of the students sitting in the classroom are useful for analyzing their learning behavior, grasping their attendance and so on during the lecture. Information about the leaning behavior or attendance of students can be used for guiding their learning activities by considering their type of participation in the class. For observing the classroom, we employ the camera attached on the notebook or tablet PC of the lecturer, because recent lectures are usually given by the lecturers with their own PCs, which are often equipped with cameras.

In the field of computer vision, there are many previous works on estimating positions of humans from camera images [8, 9]. However, the arrangement of humans in a classroom is quite different from usual scenes discussed in those previous works. Since students in a classroom sit on closely-spaced seats, many occlusions often occur among those students in their image if it is captured by the camera on the lecture's PC. However, each student keeps sitting on the same seat, and the camera on the lecture's PC can keep observing the students from their fronts throughout the whole time of the lecture. Thus, even if some students happen to be occluded by other students in their camera image at a moment of the lecture, they should appear in the image at another moment. Moreover, although all the students may not be able to be observed by the camera at the same single moment, they could be observed if the lecturer sometimes moves the PC during the lecture to change the camera angle.

Our method first recovers the 3D position of each student appearing in the camera image at each moment of the lecture from his/her face. Then the position and the orientation of the floor of the classroom in relation to the camera at each moment is estimated based on the distribution of the recovered 3D positions of the students by introducing the geometric constraint that the students sit on the seats arranged on the plane parallel to the floor. The seat location of each student is obtained by estimating the seat arrangement on the floor based on the geometric constraint that the 2D positions of the students on the floor are arranged in a grid pattern. In the remainder of this article, we will describe our method above in Sect. 2 and experimental results in Sect. 3, before giving conclusions in Sect. 4.

2 Estimating Seat Locations of Students

2.1 Recovering 3D Facial Positions

Our method first estimates the 3D positions of the students in relation to the camera from the appearance of their faces in the camera image (see Fig. 1). We detect faces in the image before extracting the eyes as well as estimating the 3D orientation for each detected face by facial image processing. For the k-th face detected in the image, let $\boldsymbol{s}_k^C$ denote the 3D position of the midpoint between the eyes in the camera-centered coordinate system with the origin at the optical center, the z axis along with the optical axis and the x, y axes parallel to the image plane. When the 2D positions of the left eye

and the right eye extracted for the k-th face in the image are represented by $\boldsymbol{l}_k^I$ and $\boldsymbol{r}_k^I$ respectively, $\boldsymbol{s}_k^C$ can be obtained from the following equations:

$$\boldsymbol{l}_k^I = \lambda A\left(R_k \boldsymbol{l}_k^F + \boldsymbol{s}_k^C\right) \tag{1}$$

$$\boldsymbol{r}_k^I = \lambda A\left(R_k \boldsymbol{r}_k^F + \boldsymbol{s}_k^C\right) \tag{2}$$

where $\boldsymbol{l}_k^F$ and $\boldsymbol{r}_k^F$ denote the actual 3D positions of the left eye and the right eye of the k-th face in its face-centered coordinate system with the origin at the midpoint between the eyes and the x, y, z axes rightward, upward and forward of the face. The coordinates of these 3D positions are given as $\boldsymbol{l}_k^F = (-d, 0, 0)^T$ and $\boldsymbol{r}_k^F = (d, 0, 0)^T$, where d denotes the intraocular distance. Matrix A, which consists of the inner camera parameters, represents the process of optical projection together with scaling parameter λ. Matrix R_i denotes the pose of the face in the camera-centered coordinate system, and is given by the 3D orientation of the face obtained by the facial image processing described above.

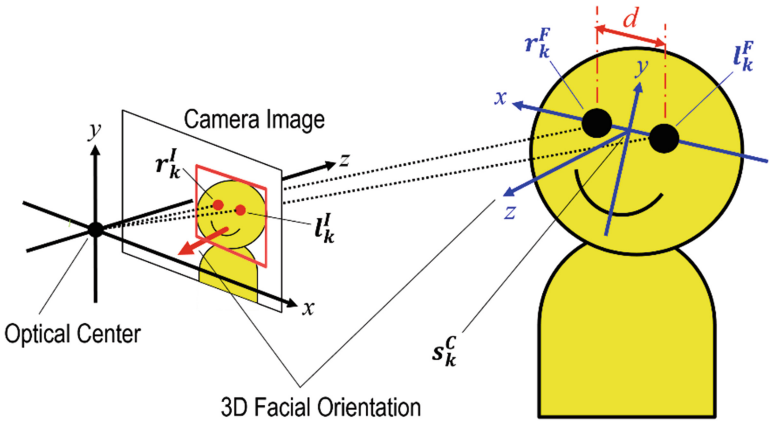


Fig. 1. Recovery of the 3D facial position relative to the camera

2.2 Estimating 2D Positions on the Floor

In order to further obtain the 2D position of the student with the k-th face on the floor of the classroom from $\boldsymbol{s}_k^C$, we need to know the position and the orientation of the floor in the camera-centered coordinate system. However, these position and orientation are unknown, because the camera is attached on the PC of the lecturer and the PC is moved so that all the students are captured by the camera. Since the students sit on the seats arranged on the floor, the 3D positions of their faces should be distributed on the same 3D plane parallel to the floor, if we neglect the difference in their sitting heights. Thus, we estimate the position and the orientation of this plane P instead of the floor by fitting a 3D plane with $\boldsymbol{s}_k^C$ for all the faces in the image.

Let us represent a 3D plane in the camera-centered coordinate system as follows:

$$\pi(\boldsymbol{x}; \alpha, \beta, \gamma, \delta) = \alpha x + \beta y + \gamma z + \delta = 0 \tag{3}$$

where $\boldsymbol{x} = (x, y, z)^T$. The normal vector of the plane is represented by $\boldsymbol{n} = (\alpha, \beta, \gamma)^T$ and the position of the plane among all the planes with the same normal vector $\boldsymbol{n}$ is specified by δ. In order for $\boldsymbol{s_k^C}$ for all the faces detected in the image to be on this plane, the following constraint needs to be satisfied:

$$E_P(\alpha,\ \beta,\ \gamma,\ \delta) \equiv \sum_{k=1}^{N} \pi\left(\boldsymbol{s_k^C};\ \alpha,\ \beta,\ \gamma,\ \delta\right)\ = 0 \tag{4}$$

where N is the number of the faces detected in the image. The values of α, β, γ and δ satisfying Eq. (4) are obtained by minimizing E_P with these variables.

The 2D position of the student with the k-th face on the floor is denoted by $\boldsymbol{s_k^P}$. Since the floor and the plane P are parallel to each other and both perpendicular to $\boldsymbol{n}$, we represent $\boldsymbol{s_k^P}$ by the plane-centered coordinate system with the x, y axes orthogonal to each other and both perpendicular to $\boldsymbol{n}$. The unit vectors of these x, y axes are denoted by $\boldsymbol{p_x}$ and $\boldsymbol{p_y}$ respectively. We set $\boldsymbol{p_x} = (1, 0, 0)^T$ so that it coincides with the x axis of the camera-centered coordinate system and $\boldsymbol{p_y} = \hat{\boldsymbol{n}} \times \boldsymbol{p_x}$, where $\hat{\boldsymbol{n}} = \boldsymbol{n}/\|\boldsymbol{n}\|$. In this plane-centered coordinate system, $\boldsymbol{s_k^P}$ is given as follows:

$$\boldsymbol{s_k^P} = \begin{pmatrix} \boldsymbol{p_x^T} \\ \boldsymbol{p_y^T} \end{pmatrix} \boldsymbol{s_k^C} \tag{5}$$

2.3 Obtaining Seat Location of Each Student

In order to know the seat location (the row and column of the seat) of each student in a grid arrangement of the seats in the classroom, we further need to obtain the position and the orientation of the grid arrangement on the floor. Let $\boldsymbol{s_k^S} = (i_k, j_k)^T$ denote the seat location of the student with the k-th face in the image, where i_k and j_k are the numbers of the column and the row of the seat counted from the seat at a corner of the grid seat arrangement, and the seat at the corner is represented as $(0, 0)^T$. When we represent the 2D position of this corner and the orientation of the grid seat arrangement by $\boldsymbol{o^P}$ and R^P respectively, the following equation should be satisfied for $\boldsymbol{s_k^P}$ and $\boldsymbol{s_k^S}$ for all the faces in the image:

$$\sigma\left(\boldsymbol{s_1^S}, .., \boldsymbol{s_N^S};\ \boldsymbol{o^P},\ R^P\right) \equiv \sum_{k=1}^{N} \left\| R^P \begin{pmatrix} wi_k \\ hj_k \end{pmatrix} + \boldsymbol{o^P} - \boldsymbol{s_k^P} \right\|^2 = 0 \tag{6}$$

where w and h are constant values for the intervals between adjacent columns and adjacent rows of the grid seat arrangement. Moreover, the following constraint should also be satisfied because different students do not sit on the same seat:

$$\delta\left(\boldsymbol{s}_1^S, .., \boldsymbol{s}_N^S\right) \equiv \sum_{\substack{k,\, l \,=\, 1 \\ k \,\neq\, l}}^{N} \left\|\boldsymbol{s}_k^S - \boldsymbol{s}_l^S\right\|^2 > 0 \tag{7}$$

For finding seat locations of all the students with the faces in the image, we minimize the following function for $\boldsymbol{s}_1^S, .., \boldsymbol{s}_N^S$ together with $\boldsymbol{o}^P$ and R^P:

$$E_S\left(\boldsymbol{s}_1^S, .., \boldsymbol{s}_N^S; \boldsymbol{o}^P, R^P\right) \equiv \sigma\left(\boldsymbol{s}_1^S, .., \boldsymbol{s}_N^S; \boldsymbol{o}^P, R^P\right) - \delta\left(\boldsymbol{s}_1^S, .., \boldsymbol{s}_N^S\right) \tag{8}$$

2.4 Merging Results in Different Frames

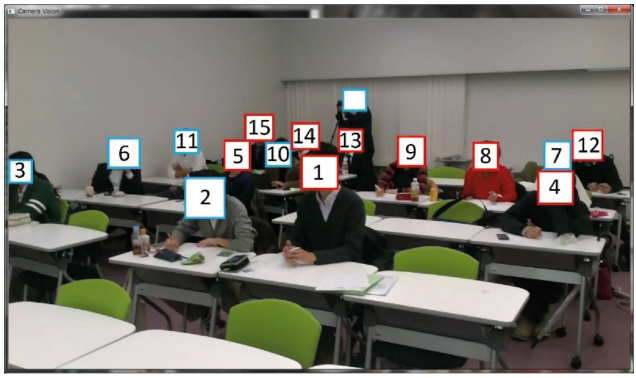
In order to obtain the seat locations for all the students whose faces can be captured by the camera at any moment during the lecture, we need to merge the seat locations obtained for the images of all the frames. Since $\boldsymbol{o}^P$ could be different for the images of different frames, we need to shift the seat locations obtained for the faces detected in the images of different frames before overlapping them.

Among all the pair of faces detected respectively in the t-th frame and the $(t + 1)$-th frame, we first find the pairs recognized as the same face by face recognition using facial image processing. Then we overlap the seat locations obtained for those frames by shifting the seat locations obtained for the $(t + 1)$-th frame so that the same seat location is assigned to the pair of the faces that are given the best similarity by the face recognition. In this overlapping, other pairs of faces recognized as the same face may be given different seat location. Currently, we simply give higher priority to the seat location obtained in the latter frame, although we may need to employ more sophisticated conflict resolution strategy in near future.

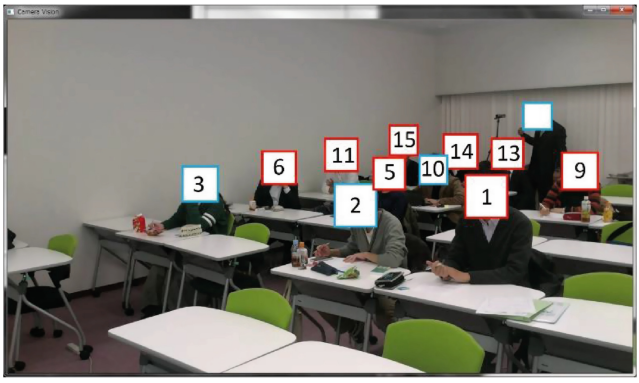
3 Experimental Results

We evaluated our method described above using the camera images of the students in a seminar supervised by one of the authors in his university. We employed OKAO Vision of OMRON Corporation for facial image processing include face detection, eye extraction, facial orientation estimation and face recognition, OpenCV for the other image processing, and the Powell method [10] for minimizing functions E_P and E_S. The value of d is set as 63 mm based on the data of the standard human face.

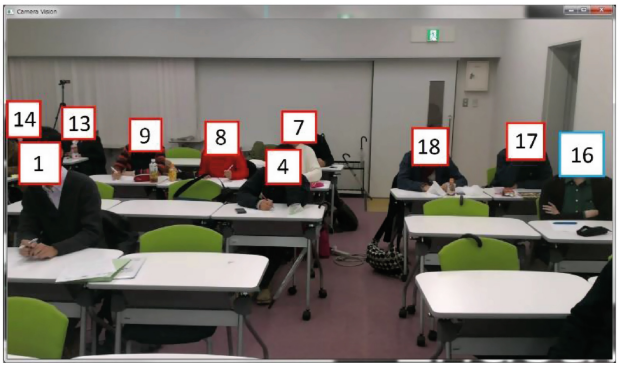
Figure 2 are sample image frames of the camera at the moments when it observes the central area, the right side and the left side of the classroom. These images are captured by the same camera oriented in different directions from around the left corner in the front side of the classroom. The numbered squares in red and blue in each image are the faces detected in that frame and the other frames. Figure 3(a)–(c) are the results of estimating the seat locations of the students whose faces are extracted in Fig. 2. Figure 3(d) is the overall seat locations obtained by merging (a)–(c). This result is obtained by consecutively overlapping the partial seat locations of (c) after overlapping



(a) Facial regions extracted from an image of the central area of the classroom

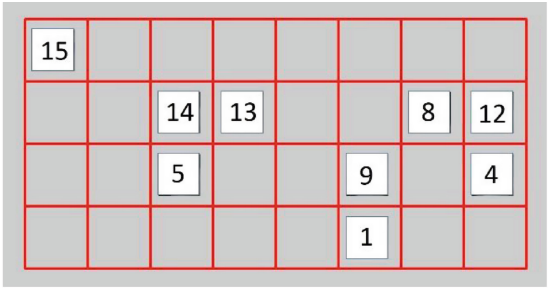


(b) Facial regions extracted from an image of the right side of the classroom

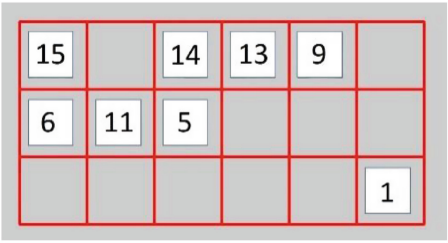


(c) Facial regions extracted from an image of the left side of the classroom

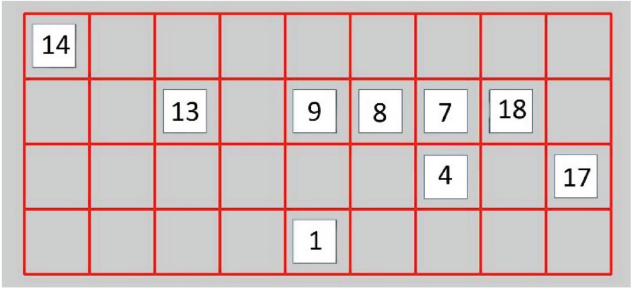
Fig. 2. Sample results for extracting facial regions (represented by the numbered squares)



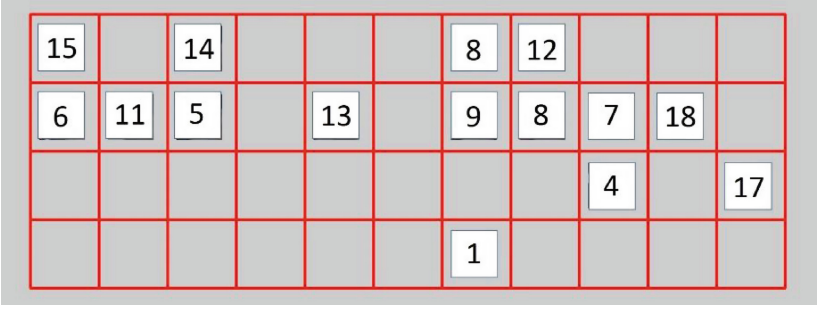
(a) Partial seat locations estimated for the facial regions in Fig.2 (a)



(b) Partial seat locations estimated for the facial regions in Fig.2 (b)



(c) Partial seat locations estimated for the facial regions in Fig.2 (c)



(d) Total seat locations after overlapping (a), (b) and (c), consecutively

Fig. 3. Resultant seat locations estimated for the facial regions in Fig. 2

the seat locations of (b) to (a). This result roughly reflects the actual positional relations among the seat locations of the students, although the rows and the columns for the seat locations of several students include errors within a single row and a column. These errors are mainly caused by the error in the process of recovering the 3D position of each students in the camera-centered coordinate system from their face detected in each image frame. As described in Sect. 2.1, this process employs the 3D orientation of the face together with the 2D positions of the eyes, which are obtained by facial image processing. However, the result of this facial image processing often includes some errors especially for small facial regions. We need to consider additional constraints to cope with this error in one of our future steps.

4 Conclusions

We proposed a method to estimate the position of each student in a classroom from the images of a camera observing the classroom, assuming that the camera is attached on the PC used for the lecture by the lecturer. First, the 3D position of each student in the camera-centered coordinate system is recovered from his/her face detected in the image using the 3D orientation of the face and the 2D positions of the eyes extracted from the detected face by facial image processing together with the knowledge of the intraocular distance of the standard human face as the clues for the recovery. Then the position and the orientation of the floor in the camera-centered coordinate system are estimated by fitting a plane to the recovered 3D positions of all the detected faces, because all the students are sitting on the seats on the same floor of the classroom. The 2D position of each student on the floor is obtained by projecting the recovered 3D position onto the estimated plane. Finally, the position and the orientation of the grid arrangement of the seats on the floor are estimated from the obtained 2D positions of all the students detected in the image. The seat location of each student is obtained by specifying the row and the column of the seat in the grid seat arrangement with the estimated position and orientation for his/her 2D position on the floor. Overall seat locations of all the students in the classroom are obtained by overlapping the local seat locations in the results for different image frames after finding the correspondence between the faces detected in those different frames based on face recognition.

We evaluated our method by the experiment using video images of a seminar. We confirmed from the result that our method can estimate rough positional relations among the seat locations of the students in the classroom, as far as their faces can be detected. However, the estimated seat locations include some errors in their rows and the columns in the grid arrangement of the seats, although the amount of errors is just within a single row and a column. Since these errors are caused by the error of facial image processing for estimating the 3D orientation of the detected face and the 2D position of the eyes by facial image processing, it is possible to assume that the faces detected in the image are oriented toward the camera. It is also useful to introduce additional geometric constraints to cope with this error. Possible constraints are the invariance of the seat location of the students and continuity of the position and orientation of the camera. These are the possible future steps of this work.

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, vol. 2774. Springer, Heidelberg (2003)
2. https://www.coursera.org/
3. https://www.edx.org/
4. 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)
5. 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), pp. 781–784 (2004)
6. Nishiguchi, S., Kameda, Y., Kakusho, K., Minoh, M.: Automatic video recording of lecture considering variety of motion and equability of scale for observing students. J. Adv. Comput. Intell. Intell. Inf. **8**(2), 180–188 (2004)
7. Wulff, B., Rolf, R.: OpenTrack – Automated Camera Control for Lecture Recording. In: IEEE International Symposium on Multimedia (ISM), pp. 549–552 (2011)
8. Turaga, P.T., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: a survey. IEEE Trans. Circ. Syst. Video Technol. **18**(11), 1473–1488 (2008)
9. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. **43**(3), 16 (2011)
10. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes, 3rd edn. Cambridge University Press, New York (2007)