



Development of Privacy Protection Monitoring Systems Using Skeleton Models and Their Evaluation on the Viewpoint of FUBEN-EKI

Hisashi Handa^{1,2(✉)}, Shingo Ando³, Tatsuhiko Ichikawa¹, Riku Yamamoto¹,
and Miyu Otani¹

¹ Faculty of Informatics,
Kindai University, Higashi-Osaka 577-8502, Japan
handa@info.kindai.ac.jp

² Research Institute for Science and Technology,
Kindai University, Higashi-Osaka, Japan

³ Graduate School of Science and Engineering Research,
Kindai University, Higashi-Osaka, Japan
<https://www.info.kindai.ac.jp/handa>

Abstract. In this paper, we propose a monitoring system using the pose estimation, which takes account of the protection of privacy. In order to cover a wider area, Region Correspondence Mechanism which associates skeletons of multiple cameras is introduced. In this paper, we examined the accuracy of association of the Region Correspondence Mechanism and verified the confidentiality of the skeleton model. Furthermore, qualitative consideration was made on the benefit of the inconvenience of the proposed system, i.e., FUBEN-EKI. We concluded that there are FUBEN-EKI such as “Enhance awareness”, “Understand systems”, and “Feel at ease”.

Keywords: Pose estimation · Region Correspondence Mechanism · Privacy protected monitoring system · FUBEN-EKI

1 Introduction

In this study, skeleton models are constituted using pose estimation [1, 2], and a monitoring system is constituted by depicting a skeleton in a background image. By constructing a monitoring system with a skeleton model, it is possible to infer the existence of a person and the action taken by the person, but it becomes difficult to identify a person. Hence, this monitoring system could be a privacy-protected system. Such a privacy protection monitoring system is useful in the fields of medical care, nursing care, and education. In these fields, monitors are often located in places that are seen by third parties. Since the developed privacy protection surveillance system is implemented using machine learning,

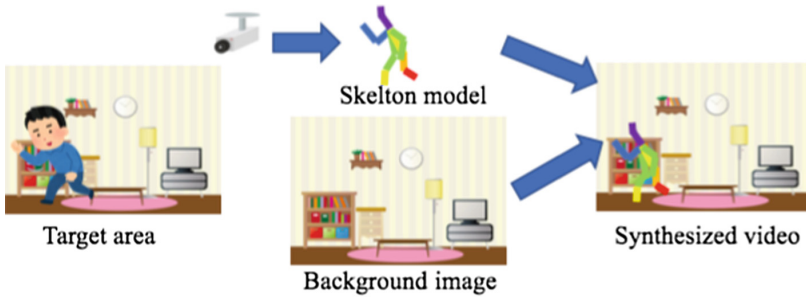


Fig. 1. A depiction of the developed systems

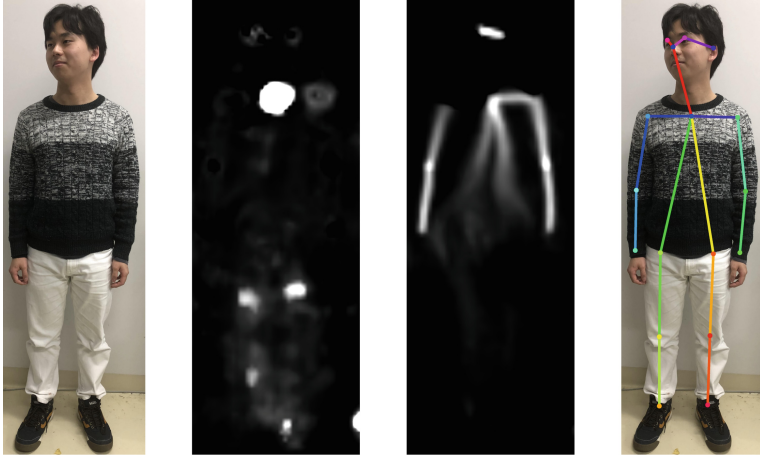
it is possible to display not only the skeleton model but also the full name of the worker in the text. From the viewpoint of FUBENE-EKI, this paper evaluates the effectiveness of the display of this monitoring system using skeleton models rather than text.

Figure 1 shows a depiction of the developed system: First, the network camera will take a video file the target area. Recorded videos are saved in the NAS. Moreover, the Pose estimation algorithms by Cao et al. is used to generate the skeleton models by applying to each frame in the recorded video file. Then, the system draws the generated skeleton model in the background image. The background image is taken in advance so that it is different with the one in live-streaming video. By sequentially displaying the synthesized images, the system can create a moving image in which persons are replaced with skeleton models.

2 Pose Estimation

This section introduces Cao's Pose Estimation Algorithms [2]. The method can detect poses in the image with CNN. There are two major approaches for the pose estimation: bottom-up and top-down approaches. In the bottom-up approach of pose estimation using images, it is difficult to join estimated human's parts [3]. On the other hand, in the top-down approach, there is a problem that the calculation cost increased in proportion to the number of people. Cao's Pose Estimation Algorithm is a bottom-up method that addresses these problems and can perform pose estimation with one inference. It enables high-speed processing in real time. Even when multiple people are shown in the input image, it can detect feature points of each person.

First, using the input image, Predict Part Confidence Maps are generated by detecting the position of the part of humans. Next, the Part Affinity Fields are generated, which includes the position information of each part and the skeleton direction information, encoding the degree of relevance between the parts. Based on Part Confidence Maps and Part Affinity Fields, a skeleton model is yielded by bipartite matching of plausible combinations of parts. Part Confidence Maps is a network that predicts position information of each part of the body in a heat



(a) Input image (b) Part Confidence Map (c) PAFs (d) Pose-estimated image

Fig. 2. An example of pose estimation

map. The Part Confidence Maps are generated for each part of a human. Part Affinity Fields represents the possibility of connecting each part. It is a network that predicts a vector map.

The error functions of Part Affinity Field and Part Confidence Maps are as follows:

$$f_{\mathbf{S}}^t = \sum_{j=1}^J \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot \|\mathbf{S}_j^t(\mathbf{p}) - \mathbf{S}_j^*(\mathbf{p})\|_2^2, \quad (1)$$

$$f_{\mathbf{L}}^t = \sum_{c=1}^C \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot \|\mathbf{L}_c^t(\mathbf{p}) - \mathbf{L}_c^*(\mathbf{p})\|_2^2, \quad (2)$$

Figure 2(b) represents Confidence Map: It shows coordinates with a high probability of neck positions. Figure 2(c) indicates Part Affinity Fields: It estimates the connections of each part. As a consequence of the pose estimation, 18 parts are detected as shown in Fig. 2(d).

3 Matching Persons Between Two Cameras

A diagram of a naive matching method of this research is shown in Fig. 3: For the cameras A and B, the skeleton information of the motion picture is obtained by the pose estimation model. We associate the positions of persons in the two video files with the region correspondence method.

3.1 Region Correspondence Between Two Cameras

As shown in Fig. 3, a model for region correspondence will be described. First, we prepare video files by two cameras with a set time. Then, for each video file,

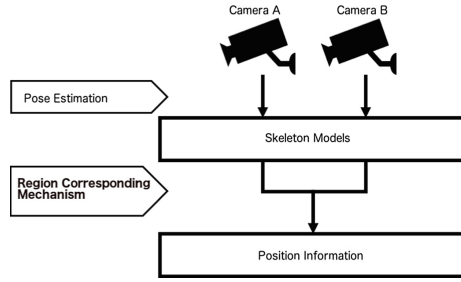


Fig. 3. Diagram of matching method

the positions of the neck, chest, and waist are estimated by the pose estimation method in the previous section. Since the pose estimation method can yield the position of the neck, the left-waist, and the right-waist, the following calculations to estimate the position of the chest and waist is adopted: The position of the waist was the coordinates of the middle point between the left-waist, and the right-waist. The coordinates of the point internally dividing the line segment connecting the neck and the waist into 1: 2 was used as the chest position. The reason for using the neck, chest, and waist is that the upper body is less incorrect estimation in pose estimation than the lower body.

The frame rate of pose estimation is 30 fps. The position of the neck, chest, and waist is given by the position in 19201080 pixels. The position on the pixel is sampled every 20 pixels. Each 2020 region is numbered sequentially from the upper left area of the image. As a consequence of the sampling, the position of the neck, chest, and waist is given by the region numbers.

The model is composed of maps of (the region number, the list of the number of another camera). For each camera, such a model is constituted. A list is used because there are multiple estimated regions in another camera for a certain region. Such multiple correspondences are caused by the height of a person and the time shift between cameras.

Region correspondence between two Cameras shown in Fig. 3 is described. As in Fig. 4, we employ video files that start at the same time, where two people are walking freely. First, the coordinates of the neck of two persons are estimated and converted into the region numbers. Figure 5 shows the region numbers of the two necks at the same time in cameras A and B.

As shown in Fig. 6(a), two persons are assumed to be person 1 and person 2. The list of region number of person 1 in the camera B is searched using the model of the camera A. Suppose that there are two persons in camera B: person I and person II. Let $d_{1,I}$ be the distance between the region number of person I and the list of the region numbers of person 1 in the camera B, where distance metrics is used as the Manhattan distance in the region numbers in video image. As in Fig. 6(b), the correspondence of persons 1, 2 to persons I,



Fig. 4. An example of the video file: two persons walk around.

time	Camera A				Camera B			
	ID_a1	a1_id	ID_a2	a2_id	ID_b1	b1_id	ID_b2	b2_id
1	no1	a2084	no2	a2065	no1	b1386	no2	b1208

Fig. 5. Region numbers of two necks at the same time

II is made by comparison of the sum of distances: $d_{1,I} + d_{2,II} + d_{I,1} + d_{II,2}$ and $d_{1,II} + d_{2,I} + d_{II,1} + d_{I,2}$. If the sum $d_{1,I} + d_{2,II} + d_{I,1} + d_{II,2}$ is smaller than another sum, the method detects that person 1 in camera A corresponds the person I in camera B. Otherwise, the method detects that person 1 in camera A corresponds the person II in camera B.

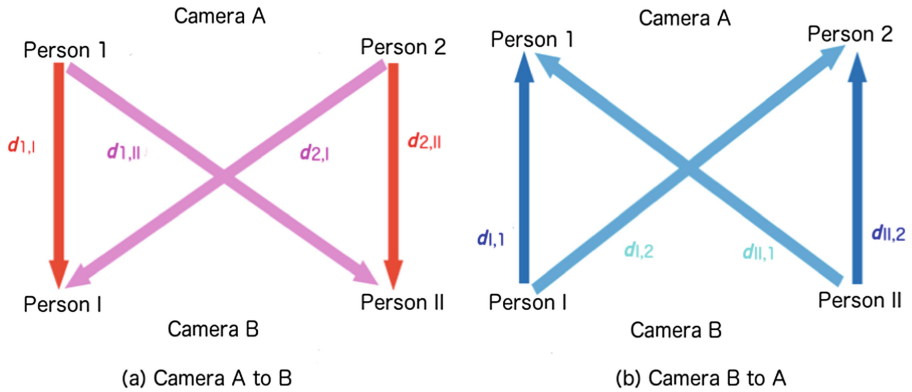


Fig. 6. Region correspondence

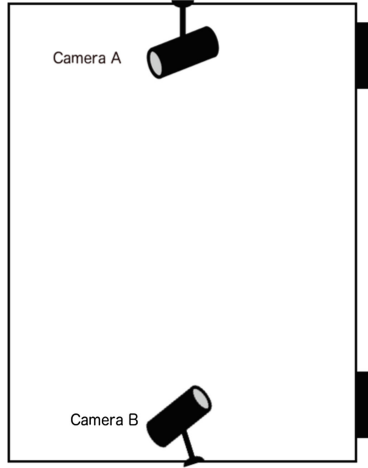


Fig. 7. Placement of cameras

4 Experiments

4.1 Experiential Configuration

We used a wide-angle network camera for the experiment. The specifications are summarized as follows:

- A fisheye lens that can display a wide range of about 180° horizontally
- Resolution: 1920 1080 pixel
- 15 fps

Wide-angle cameras are difficult to make low distortion but can take images with deep depth of field. Since the wide-angle camera has a wide imaging range, it leads to cost reduction of the entire system. This is the reason for using a wide angle camera.

Figure 7 shows the placement of the camera. Two wide-angle cameras are installed in the center of the short side wall of the room. Each camera stands up with one pole erected. Suppose that two cameras be cameras A and B, respectively. Such a camera placement can photograph the entire room and the whole body by looking down at a person and photographing it. It prevents several people from overlapping on the image. Figure 8 shows an example of images at the same time from each camera.

Table 1 shows three models used in this paper. Columns “A to B” and “B to A” stand for the number of the region numbers in the maps for each camera. Bigger models are expected to have better performance. Model 1 is a model generated using the region numbers at the neck and waist positions. Model 2 and 3 are of the region numbers at the neck and chest, and at the neck, chest, and waist, respectively.

Table 1. Models

Model	A to B	B to A
1: neck, and waist	1412	1532
2: neck, and chest	1235	1447
3: neck, chest, and waist	1479	1700



(a) snapshot of Camera A



(b) snapshot of Camera B

Fig. 8. An example of images at the same time from each camera

4.2 Experiments on Person Tracking

The accuracy of estimating the data of 1171 necks for validation, according to the model of Table 1 is shown in the Table 2. The accuracy means that the ratio of correct detection over the number of the region numbers in the model, where the correct detection means that two persons in each camera are correctly estimated. The number of regional numbers indicates the number of times the region number of the validation set is found in the region number in the model.

We consider the relationship between the number of region numbers in Table 1 in Sect. 3.1 and the part used for the model. Model 2 using the neck and chest data has fewer searchable region numbers than model 1 using neck and waist data: 177 in “A to B”, and 398 in “B to A.” As a result, the number of estimable regions is 56 less.

Table 2. Accuracy of each model

Model	Accuracy	No. region numbers	No. uncorresponding region numbers
1: neck, and waist	98.96%	1158	13
2: neck, and chest	99.31%	1165	6
3: neck, chest, and waist	99.74%	1171	0

However, four of the 12 misdirected estimates in Model 1 are correctly estimated in Model 2, The number of incorrect estimates has been reduced to 8, and



Fig. 9. Scatter plots of distances (Model 1)

the accuracy is increased by 0.35 points. Also, in Model 3 including the neck, waist, and chest, it is possible to estimate all the validation data. From these results, it can be seen that even if the number of the region numbers increases and the number that can be estimated increases, the estimation accuracy is not necessarily improved. This is attributed to the difference in height between the parts used to generate the model, i.e. the waist, and the parts used in the validation, i.e., the neck. This is because, in one camera, in the case where the waist of a person is located in a certain region number, and in the case where the neck is positioned in that region number, the region number in the other camera is totally different.

To find out the cause of the incorrect estimations, Fig. 9 shows the experimental results for 900 frames of a verification movie of each model. The green square “correct” indicates that the correspondence of the persons to the walking of the two persons in the video frame is successful. On the other hand, the red circle “miss” denotes incorrect correspondence. The x-axis represents the Manhattan distance of the region number in camera A, and the y-axis represents the one in camera B. It is often the lower left area of the graph that makes incorrect estimations in any model. This means that the Manhattan distance in the cameras A and B is 10 or less for both cameras. That is, it shows that the two persons are close. The number of incorrect estimations of Model 1, 2, and 3 is 9, 6, and 3, respectively (Figs. 10 and 11).

Figure 12 is a scatter plot of performing region correspondence that two people pass each other at the viewpoint of a camera. The length of the video file is nine seconds. Model 3 is used for this scatter plot. This model correctly estimates the correspondence of people.

The images of cameras A and B at the time when the two Manhattan distances at Camera A are 4 and at Camera B is 30 are shown in Fig. 13. When the

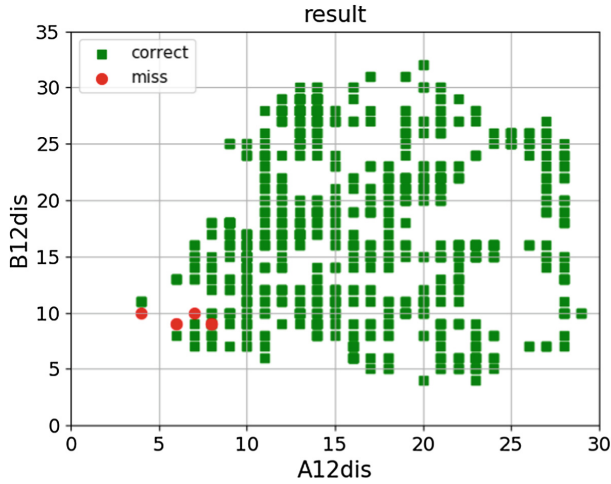


Fig. 10. Scater plots of distances (Model 2)

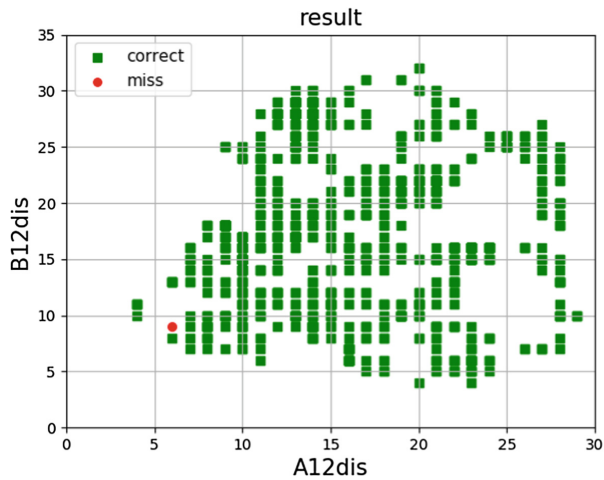


Fig. 11. Scater plots of distances (Model 3)

distance between the two Manhattan at Camera A was short, Camera B showed Manhattan distance and it was reflected. As a result, it is considered that a large distance difference occurred in the area to be estimated, and it was estimated correctly.

4.3 Experiments of Privacy Protection Monitoring Systems

As mentioned previously, by adding a skeleton model to the background image, we construct a monitoring system with privacy protection. Such overwriting is

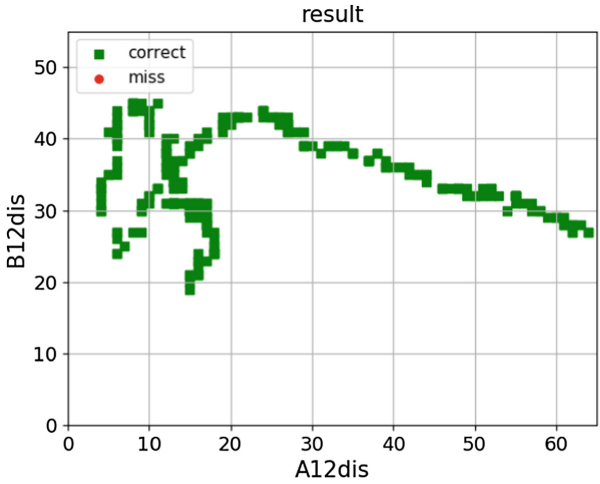


Fig. 12. Scatter plot of that two people pass each other

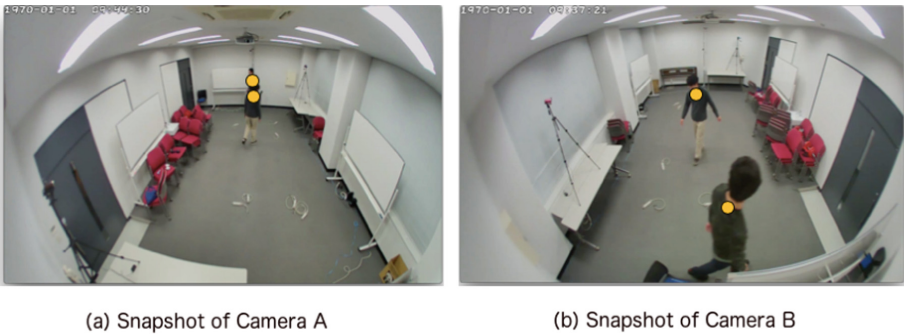


Fig. 13. Viewpoint at camera A and B

performed for each frame of the video file. By concatenating the modified frames, it is possible to realize a monitoring system in which the skeleton model runs. Note that items and desks in images cannot move because only one background image is used. By linking with the region correspondence system in the previous section, monitoring can be performed while always tracking people in a wide area.

In this section, we examine the effectiveness of privacy protection by using the skeleton model. We ask four subjects to do muscle training. We take a video the muscle training. Then, the skeleton model is constituted by using the proposed system. Figure 14 shows snapshots of original video doing muscle training and corresponding skeleton video for the monitoring system. Obviously, the skeleton model contributes to privacy protection.

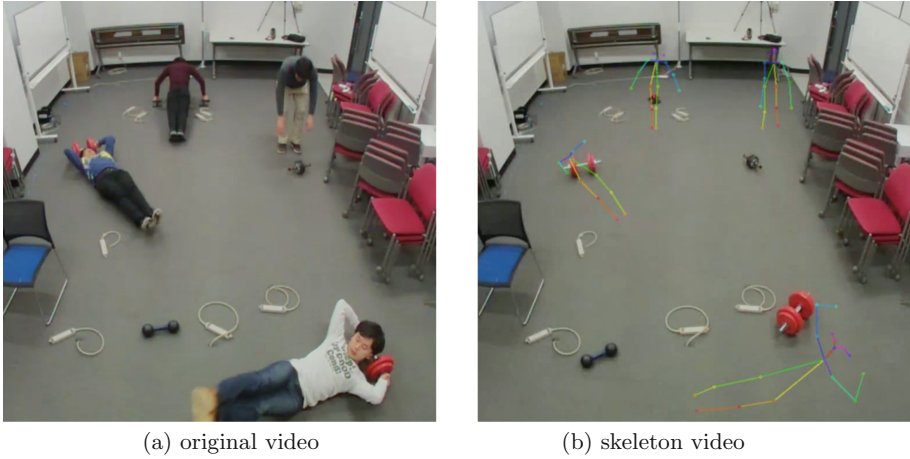


Fig. 14. Snapshots of original video doing muscle training and its skeleton video for the monitoring system

By watching skeleton video, the other nine subjects will guess who each of the four skeleton models is. They know the four subjects well but they do not know that the four subjects were doing muscle training. The nine subjects are told that the four subjects are being photographed, and one person is selected for each skeleton model from the four subjects. The length of the skeleton video is about 30 s. Table 3 shows the confusion matrix for the guess. Two of the four subjects, i.e., S, I, are shorter than the others. Hence, the nine subjects seem to be classified if they are tall or not.

5 Qualitative Discussion from the Viewpoint of the FUBEN-EKI

This monitoring system is considering use in nursing care facilities. If the latest machine learning methods are used, we could generate text information about when and who cares to someone when analyzing the video. The proposed systems introduced inconvenience, i.e., FUBEN, of “Less information”, “Time consumption”, and “Continuity.” Such FUBENs not only contribute to privacy protection but also the benefit, i.e., FUBE-EKI, such as “Enhance awareness” and “Understand systems.”

Person estimation from the skeleton model takes a very long time. But if you know the behavior and gesture of that person, its estimation will be easy. Furthermore, instead of converting actions into texts, by presenting them in an analog skeleton model, if the carer is ramped up or an accident occurs, it is easy to detect such abnormalities. That is, it also brings FUBEN-EKI “Feel at ease”.

Table 3. Confusion matrix

	K	M	S	I
K	5	3	1	0
M	3	4	1	1
S	0	1	4	3
I	1	1	3	4

6 Conclusions

In this paper, we have constructed a monitoring system taking care of privacy protection using pose estimation. In order to cover a wide area using multiple cameras, we introduced a method of tracking people by the region correspondence mechanism. By combining Pose estimation and region correspondence system, it is possible to construct a system to monitor while protecting privacy. Section 4.2 examined the accuracy of the region correspondence mechanism. Although it was the correspondence between two cameras and two people, it was found that correspondence can be done with good precision. The confidentiality of the skeleton model was verified in Sect. 4.3. The benefits of the inconvenience, i.e., FUBEN-EKI, related to the proposed system were discussed in Sect. 5. “Less information”, “Time consumption”, and “Continuity” are introduced in the proposed system. As a result, FUBEN-EKI such as “Enhance awareness”, “Understand systems”, and “Feel at ease” are obtained.

Future works are summarized as follows: First, it is necessary to apply the proposed system to multiple cameras, multiple people environments, and to examine the tracking performance. We are considering introducing it to an actual care facility and conducting demonstration experiments.

References

1. Pishchulin, L., et al.: Deepcut: joint subset partition and labeling for multi person pose estimation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
2. Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations (ICLR) (2015)
4. Takahashi K.: Camera Calibration Based on Mirror Reflections (2018). <http://hdl.handle.net/2433/232407>