 doi:10.20944/preprints202003.0376.v1

Article

Prevention of Unintended Appearance in Photos Based on Human Behaviors Analysis

Yuhi Kaihoko [1,*], Phan Xuan Tan [2] and Eiji Kamioka [1,*]

[1] Graduate School of Engineering and Science, Shibaura Institute of Technology; {ma18028, kamioka}@shibaura-it.ac.jp
[2] Department of Information and Communications Engineering, Shibaura Institute of Technology; tanpx@shibaura-it.ac.jp
* Correspondence: kamioka@shibaura-it.ac.jp (E.K.); tanpx@shibaura-it.ac.jp (P.X.T)

Abstract: Many people can take photos with smartphones and easily post photos via SNS (Social Network Services). This has caused a social problem that unintended appearance in photos may threaten the privacy of photographed persons. For this issue, numerous studies have already been introduced to prevent the unintended appearance in photos from the photographer's side, but only a few methods tackled this from the photographed person's side. Therefore, we considered calling attention to a situation that a photo-taking behavior by a photographer can be automatically detected by using a wearable camera worn by a photographed person. In this paper, we propose an approach to detect photo-taking behaviors in video data taken from the wearable camera, analyzing specific human skeleton information. OpenPose is utilized to obtain the human's skeleton information and the time-series data are analyzed. In addition, we compare two similar behaviors which are photo-taking behaviors and net-surfing behaviors. These video data are distinguished by DP matching and cross-validation. Finally, it is concluded that the detection accuracy of photo-taking behaviors is about 92.5%, which is satisfactory enough for this research purpose.

Keywords: Photo-taking Behaviors; Behaviors detection; Smartphone; OpenPose; Skeleton.

1. Introduction

For years, the smartphone has increasingly become one of the indispensable personal devices, by which people can easily take photos to record every desirable moment just by a simple click. According to the Global Digital 2019 report, the number of people around the world who use a mobile phone account for 67 percent – more than two-thirds of the total global population [1]. In Japan, the statistics obtained from the Ministry of Internal Affairs and Communications show that the ownership rate of the smartphone is about 60.9%, and especially the rate of owners who are under 40 is over 90% in 2018 [2, 3]. This facilitates the explorations of various Social Networking Services (SNS) (e.g., Facebook, Twitter, etc.). About 3.5 billion people accounting for 45% of the global population are using SNS [1]. Therefore, people can easily take photos and post them on SNS for their purpose, leading to the so-called "unintended appearance in photos" as an increasingly social problem. This happens when someone is taking a photo in a public place, then other people may unintendedly appear in the photo which may be posted on an SNS, raising the question of privacy. To prevent such a situation, there exists numerous tools performing image processing techniques on the photos, e.g. superimposing pixelated, blurred images or putting the seal around the face in automatic or manual manners [4, 7]. Frome et al. [4] proposed privacy protection for all people in Google Street View. In order to protect the privacy of passersby who are captured by an in-vehicle camera such as Google Street View, privacy protection is applied to all persons in the image. In this method, for a given image, they run the primary and secondary face detectors in parallel. Then, the outputs are processed by using the height-to-bottom ratio, the color model ratio, secondary detector overlap, the convolution neural network output and so on. As a result, the image is determined whether blur or don't blur by the output score. This method is effective to use in public situations, e.g. uploading

 doi:10.20944/preprints202003.0376.v1

photos to the Internet and protecting all personal privacy. However, it is no meaning to blur the face of all persons who appear in the images when used in a general situation of taking pictures. In addition, the privacy risk still stays the same since this process will only be performed by photographers. In fact, there are few works in the prevention of unintended appearance in photos from photographed person side. Yamada et al. [5] proposed a method to avoid the unintended appearance in photos physically using a privacy visor that uses infrared rays. That privacy visor's shape is like a pair of glasses that are equipped with near-infrared LEDs. Lighting near-infrared LEDs cannot be recognized by human vision when the camera is aimed at the user who wears the visor. However, the user's face is obstructed by infrared rays when viewed through the camera. Thus, the photographed person can prevent the unintended appearance in photos by wearing the visor. It is stated that face detection becomes impossible by looking through the camera because of the infrared light that does not affect human vision, and it is said that it has little effect during general communication. This is a groundbreaking study that focuses on the difference between human vision and the sensitivity of camera devices. However, the photographed person must wear a privacy visor with a power supply and a power cord. This can be very burdensome for users.

In this study, we propose an approach, that is to say, detecting photo-taking behaviors by a smartphone, to prevent the abovementioned privacy issue from the photographed person's side. By leveraging a wearable camera worn by a photographed person, we believe that such behaviors can be detected. However, misdetection may occur when the likely-photographer performs a similar task with similar behavior, for example, net-surfing. Therefore, in the proposed approach, we focus on the movements of the specific human skeleton information such as arm's length and angle transition which are extracted by OpenPose [6], to identify photo-taking behaviors. More concretely, the wearable camera worn by a photographed person will continuously capture the video of surrounding people. In real-time, OpenPose extracts human skeleton information (skeleton joint coordinates). The DP (Dynamic Programming) matching between monitored data and reference data is performed to generate DP score which is then compared with the pre-defined DP threshold. The comparison results decide whether such input behaviors are photo-taking behaviors or not. The experimental results demonstrate that the proposed approach achieved an accuracy of 92%. From the evaluation experiment, the photo-taking detection accuracy is sufficient to show the effectiveness of this proposed method. The contributions of our works are described as follows:

- A novel method for photo-taking detection is proposed.
- Datasets for training and testing phases of the proposed approach are established.
- A DP threshold which plays a crucial role in our proposed method is determined.

The remainder of the paper is organized as follows: Related work is provided in section 2. Meanwhile, section 3 describes the proposed method. Performance evaluation of the proposed method is discussed in section 4, and section 5 concludes this study.

2. Related Works

Koyama et al. [7] indicated that social videos posted via SNS such as YouTube and Facebook include not only intentionally captured people (ICPs) but also non-ICPs. In order to protect the privacy of non-ICPs, the authors introduced a new system for automatically generating privacy-protected videos in real-time. Previous privacy protection systems simply blur out all the people in the video without distinguish between ICPs and non-ICPs, resulting in making an unnatural video. On the other hand, their proposed privacy protection system automatically discriminates ICPs from non-ICPs in real-time based on the spatial and temporal characteristics of the video, and then, only the non-ICPs can be localized and hidden. However, such a privacy-protected video still has some unnatural parts. In addition, since this processing is performed by the photographer, how to handle the obtained processing data depends on the morals of the photographer.

Alternatively, deep learning based approach can potentially identify photo-taking behaviors. For example, YOLO (You Only Look Once) is proposed by Redmon et al. [8]. By using Neural Network, namely Darknet, YOLO can perform image classifications and real-time object detections with high accuracy. However, to recognize behaviors of taking photos, this approach has to face

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

multiple challenges. First, enormous image data and time are required for the training process. Second, it is still difficult to recognize "overlapping objects" (e.g., a human is holding a smartphone on the right hand, etc.). Third, small objects like a smartphone that someone is holding in his/her hand in the photo cannot be detected properly. As a consequence, this approach is not suitable for photo-taking behavior detection.

In a previous study [9], we proposed methods of smartphone detection using geometric features. Accordingly, a smartphone can be detected by relying on two preconditions: (1) a photographed person (user) monitors surroundings with a wearable camera and (2) a user holds a smartphone by hands in front of his/her face. This method allows detecting the smartphone from geometric features. For instance, a straight-line detection was able to detect the smartphone's characteristic shape such as a right-angle and a parallel straight edge lines. The experimental results demonstrate that the accuracy of smartphone detection is 53.3% from the FRR-FAR curves. Thereby, a smartphone can be detected even if the smartphone is held by hands. Although the effectiveness of this proposed method was shown, the accuracy was not so high. More importantly, the final purpose of photo-taking behavior detection has not been achieved.

Regarding the behavior detection methods, Tsai et al. [10] proposed an Optical Flow based approach to detect and analyze specific human behaviors. They focused on the detection of three behaviors (Smoking, Drinking, Phoning) and others according to four states by using Optical Flow. Based on the recognized human's face, each behavior was detected with high accuracy by using Optical Flow with vector norms and color histograms. Although this approach demonstrated an effective manner in detecting and analyzing human behaviors, such detected behaviors were quite simple with limited motions on each behavior. Moreover, the behavior detection cannot be activated without initial face detection. In fact, photo-taking behaviors include more complicated motions than that. When the photographer takes photos, he/she flexibly use one hand or two hands, whereas, the smartphone direction is in landscape or portrait.

Being inspired by the success of the Optical Flow based approach in behavior detection, in [11], we proposed a method to identify photo-taking behaviors using the Optical Flow vector. In this approach, it was assumed that a photographed person (user) wears a wearable camera and monitors his/her surroundings. Based on the video data analysis, the behaviors of people who are going to take photos were early recognized. More concretely, we analyzed the characteristics of such behaviors by utilizing the Optical Flow technique referring to the movement of arms and/or hands of a human. Experimental results showed a distinctive trend in the distribution between Optical Flow norm and angle for photo-taking behaviors. In addition, the angle value of the Optical Flow norm continuously indicated 90 degrees during the behaviors ($\cos\theta = 0$), which means that the user's arm just moved vertically. The detection accuracy was about 67% across six photo-taking behaviors training datasets and six testing datasets. Eventually, we confirmed the effectiveness of the proposal, that is to say, Optical Flow based approach. However, the detection accuracy must be improved. In addition, only photo-taking behaviors were focused on without considering the discrimination from similar tasks with similar behaviors such as net-surfing.

3. Detection of photo-taking behaviors focusing on distance and angle between joints

3.1. Proposed algorithm of photo-taking behaviors detection

In this section, an algorithm to detect photo-taking behaviors which utilizes extracted human skeleton information is presented. In a general scenario, we assume that a photographed person (user) wears a wearable camera like a "Lifelog camera" and monitors surroundings. Then, based on the monitored video data, the proposed algorithm will detect someone who is in the video data is about to take a photo. The detection mechanism focuses on classifying photo-taking behaviors from net-surfing behaviors since there are similar motions of moving the arm, found in both types of behaviors. In fact, the motion to move the arm, in this case, is defined by the changes in arm's length and the angle from the view of photographed persons. Therefore, the transition of the arm's length and angle of the bending arm are crucial inputs for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

the detection mechanism. The proposed algorithm of photo-taking behavior detection is clearly described as follows:

Table 1. proposed algorithm

Proposed Algorithm
Input: *monitored video, threshold*
Output: *0*: photo-taking behaviors, *1*: net-surfing behaviors
Initiate OpenPose
Analyze the monitored video
return *25 points skeletons…*
Calculate the arm's length and angle of bending arm
(I) length of upper arm, (II) length of lower arm, (III) angle of bending arm
return *(I) ~ (III) value*
Calculate DP score
DP matching (reference data: photo-taking behaviors)
return *DP score*
If DP score <= threshold **Then**
Judged as photo-taking behaviors
return *0: photo-taking behaviors*
Else
Judged as net-surfing behaviors
return *1: net-surfing behaviors*
End if

In the following subsections, the details of each step in the proposed algorithm will be explained.

3.2. Video capture

A scenario to detect a photo-taking behavior and notify the photographed person of it is described in Fig.1. Specifically, the user wears a wearable camera and monitors surroundings. The monitored video data obtained from the camera is fed into the detection algorithm as the input for further analysis. If a photo-taking behavior is detected, a vibration signal as the output is sent to the user's smart device (e.g., smartphone).

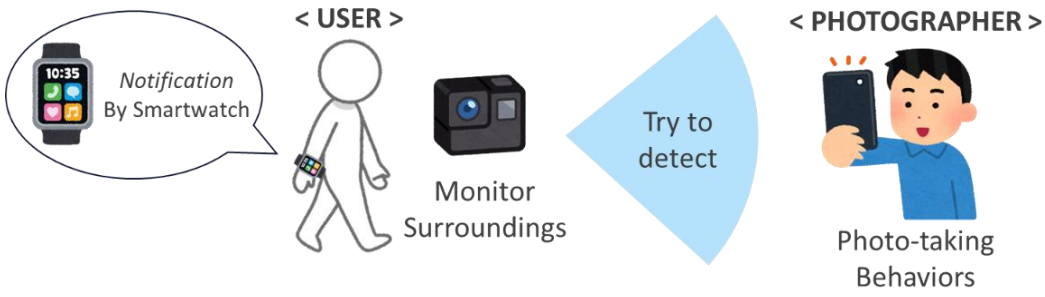


Figure 1. A scenario of photo-taking behavior detection and its notification

3.3. Extract human skeleton information

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

As mentioned in **subsection 3.1**, the main aim of the proposed approach is to detect photo-taking behaviors and classify them from net-surfing behaviors. Thus, the hand movements, which are defined by the changes in arm's length and the angle from the view of the photographed user, must be continuously tracked. For that purpose, we propose a method of behavior flows analysis that focuses on tracking the motions of a specific skeleton. More concretely, we focus on the movements of the specific human skeleton information represented by an arm's length and angle transitions to avoid detecting net-surfing behaviors as photo-taking behaviors. Indeed, when people do either taking photos or net-surfing, the movements of the left or right arm length which indicates (I) upper arm or (II) lower arm, and (III) the angle of the bending arms are likely to change. Thus, these three factors (I, II, III) are mainly focused on detection.

In order to obtain the skeleton information, OpenPose which is an open-source tool is utilized. Not only the human skeleton but also the hands skeleton and facial expression can be accurately detected by OpenPose without using a special camera. Thanks to OpenPose, the expected output can be obtained from two-dimension images that are captured by a normal camera. An example of skeleton information extracted from OpenPose is shown in Fig.2. There are 25 points connected by joint parts, establishing "BODY_25" human skeleton estimation model defined by OpenPose.

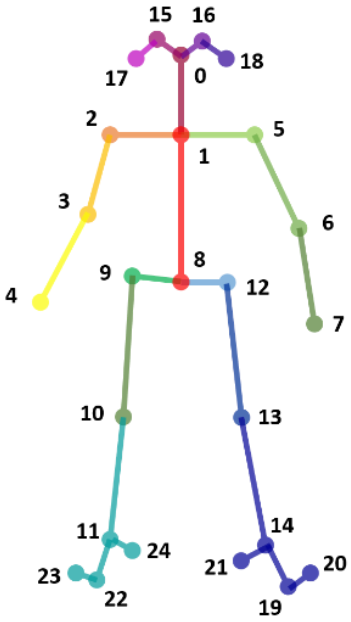


Figure 2. "BODY_25" human skeleton estimation model

OpenPose practically allows the joint coordinates to be obtained and stored as a json files for each frame. The skeleton data is therefore formed as [x, y, *confidence score*]. Here, the x and y are coordinates indicating body part locations in the input image. The *confidence score* indicates the accuracy of the coordinates calculated by OpenPose tool. As we assumed earlier, there is a potential difference in the arm's length and the angle between photo-taking and net-surfing behaviors. In this study, we focus on these parameters which are numerically calculated from joints' information. Thereby, the joints: "2, 3, 4, 5, 6, 7" as shown in Fig.2 are utilized to analyze the behaviors. These joints of arms part are briefly presented in Fig.3 and Table 2. While Fig.3 shows points and joints of the utilized body parts, Table 2 provides the correspondence between joint positions and body parts. Also, it indicates index numbers following Fig.3. Then, the upper, lower arm and angle can be calculated.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

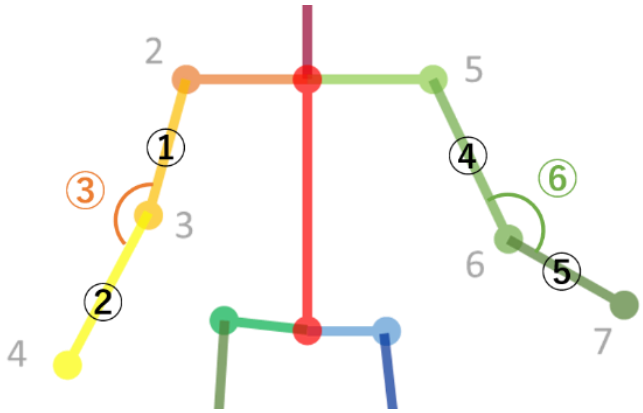


Figure 3. Image of each utilized part from OpenPose information

Table 2. Correspondence between joint position and body part

	Index factors	Joint position (keypoint)	Body part	Index number in Figure 3	Expression in this paper
Right	I	2-3	Right Upper arm	①	Length-23
	II	3-4	Right Lower arm	②	Length-34
	III	2-3-4	Angle of the bending right arm	③	Angle-234
Left	I	5-6	Left Upper arm	④	Length-56
	II	6-7	Left Lower arm	⑤	Length-67
	III	5-6-7	Angle of the bending left arm	⑥	Angle-567

According to Fig.3 and Table 2, the length and angle values are determined by using the distance between two points and inner product of coordinates obtained from OpenPose. The detailed calculation is presented in Fig.4.

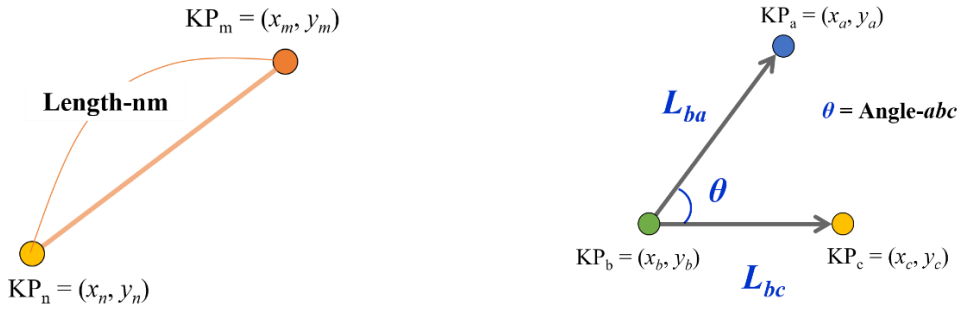


Figure 4. Calculating arm's length and angle of the bending arm. (**a**) Definition image for calculation of arm's length from two coordinates by using the distance between two points. It corresponds to ①, ②, ④, ⑤ in Table 2; (**b**) Definition image for calculating of the angle of the bending arm from three coordinates by using the inner product. It corresponds to ③, ⑥ in Table 2.

- Calculation of the arm's length (I)&(II)

According to Fig.4(a), the coordinates of a certain joint position (keypoint) form KP$_p$ with the following definition:

$$\mathrm{KP_p} = \left(x_p, y_p\right) \tag{1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

where, the p indicates a joint position (keypoint) number. The arm length can be calculated by the following equation:

$$\text{Length-nm} = \sqrt{(x_m - x_n)^2 + \left(y_m - y_n\right)^2} \tag{2}$$

In this case, "Length-nm" stands for arm's length between joint position number of n and m. By using Eq. (2), the length of the right and left upper/lower arm (indexed by ①, ②, ④, ⑤) can be properly determined.

- Calculation of the angle of bending arm (III)

The angle of bending arm θ is formed by $\text{KP}_\text{a} = (x_a, y_a)$, $\text{KP}_\text{b} = (x_b, y_b)$, $\text{KP}_\text{c} = (x_c, y_c)$ as shown in Fig.4(b). Accordingly, by using these coordinates, the angle is required after transforming to vectorization for calculating angle made by 3 points. From the above coordinates, the vectorization $\boldsymbol{L}_{pq}$ must be expressed as shown in the below:

$$\boldsymbol{L_{ba}} = \begin{pmatrix} x_a - x_b \\ y_a - y_b \end{pmatrix} = \begin{pmatrix} ba_1 \\ ba_2 \end{pmatrix}, \boldsymbol{L_{bc}} = \begin{pmatrix} x_c - x_b \\ y_c - y_b \end{pmatrix} = \begin{pmatrix} bc_1 \\ bc_2 \end{pmatrix} \tag{3}$$

By using these vectors, the angle can be calculated as expressed in equation (4):

$$\theta = \cos^{-1}\left(\frac{\boldsymbol{L_{ba}} \cdot \boldsymbol{L_{bc}}}{|\boldsymbol{L_{ba}}| \cdot |\boldsymbol{L_{bc}}|}\right) \tag{4}$$

where, $0 \le \theta \le \pi$. From this procedure, the angles of the bending right and left arms (indexed by ③, ⑥) can be calculated.

3.4. Thresholds of identifying Photo-taking behaviors comparing with Net-surfing behaviors

In this study, one of the main contributions is to clarify the threshold value of DP (Dynamic Programming) matching score, which plays an important role to decide whether a series of human hand movements is a photo-taking behavior or not. This threshold value is obtained from the point of Equal Error Rate (ERR) where False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves meet. The following subsubsections provide the definitions of DP matching, FAR, FRR and ERR in detail.

3.4.1. DP matching (Dynamic Programming matching)

DP matching is a pattern matching technique based on dynamic programming, which evaluates the similarity between two sequenced data which could consist of the different number of data points.

Initially, two patterns of sequenced data (X and Y) are expressed as follows:

$$\text{X} = x_1, x_2, \cdots, x_i, \cdots, x_I \tag{5}$$
$$\text{Y} = y_1, y_2, \cdots, y_j, \cdots, y_J \tag{6}$$

where X and Y represent a sequenced input data and the referenced sequenced data, respectively. Meanwhile, I and J indicate the number of data points of X and Y, respectively.

In addition, $d(x_i, y_j)$ expresses the distance between elements of X and Y. It will be transformed from x-y coordinates space to i-j coordinates space as follows:

$$l(i,j) = d(x_i, y_i) = \left|x_i - y_j\right| \tag{7}$$

In addition, the accumulated distance is expressed by g(i, j) in i-j coordinates space, which can be obtained by calculating the minimum DP path in an optimal distance problem. Fig.5 shows the weight for calculating optimal distance as the definition of DP path.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

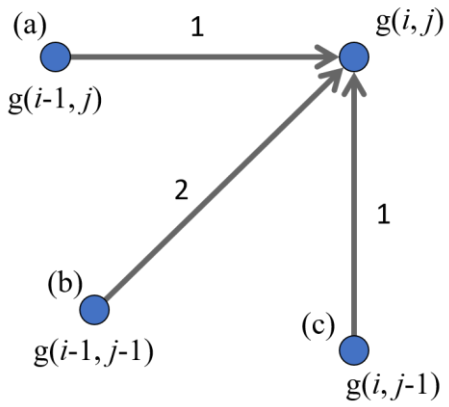


Figure 5. Definition of DP path. To calculate the accumulated distance, (a) to (c) indicates a pattern of distance in *i-j* coordinates space. Each number shown in (a) to (c) expresses each weighted score for calculating the distance by following Equation (8).

Based on Fig.5, the dissimilarity g(*i*, *j*) can be defined by:

$$g(i,\ j) = \min \begin{Bmatrix} g(i-1,\ j) + d(i,\ j) & : (a) \\ g(i-1,\ j-1) + 2d(i,\ j) & : (b) \\ g(i,\ j-1) + d(i,\ j) & : (c) \end{Bmatrix} \quad (8)$$

By using Equation (8), the accumulated distance of DP matching is expressed by g(*i*, *j*). Finally, DP matching score between X and Y is obtained by normalizing g(*i*, *j*) with the number of each data points as shown by Equation (9).

$$\text{DP score} = \frac{g(I,J)}{I+J} \quad (9)$$

The smaller the DP score is, the higher the similarity between the two data is. In this study, the utilization of DP score is two-fold. First, in the training phase, DP scores are obtained from the training datasets of the monitored human skeleton to generate the FAR and FRR, which are then plotted to determine the intersection point of EER value. As a result, the threshold value of DP score reflecting EER value is eventually determined. Second, in the testing phase, DP score which is calculated from the matching between the input skeleton information and the referenced dataset is compared with the threshold value of DP score to conclude whether the monitored hand movements are photo-taking behaviors or not.

3.4.2. "False Rejection Rate" and "False Acceptance Rate" (FRR and FAR) and the determination of the threshold

In this part, we provide the explanations on how we define FAR and FRR for the determination of EER value which is referred to as the threshold value of DP score. The terms of FAR, FRR and ERR are common in the topics of biometric security systems where:

False Acceptance Rate (FAR): the percentage of identification instances in which **unauthorized persons** are incorrectly accepted. (This is also known as False Match Rate.)

False Rejection Rate (FRR): the percentage of identification instances in which **authorized persons** are incorrectly rejected. (This is also known as False Non-Match Rate.)

In other words, FAR implies how high your system's security level is, whereas, FRR reflects the level of comfortableness of the users. In order to evaluate the operating performance of a security system, the Equal Error Rate (EER), which is also known as the Crossover Error Rate (CER), must be taken into account. It means that the system has parameters that can be turned to adjust FAR and FRR to the point where both of them are equal. Importantly, the smaller the ERR is, the better the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

performance is. In this study, the FAR is defined as the rate when net-surfing behaviors are recognized as photo-taking behaviors, whereas, FRR refers to non-detection rate of photo-taking behaviors. Specifically, the high value of FAR reflects a low performance of identification or classification of photo-taking behaviors from the others, leading to the incorrect notifications sent to the user. The higher value of FFR, on the other hand, shows the lower performance of verification or detection of photo-taking behaviors. The obtained EER is illustrated in Fig.6. In this Figure, FAR and FRR are plotted as function curves of similarity level which is defined by DP scores. The intersection point of these curves is ERR. The corresponding DP score value of ERR is the desirable threshold.

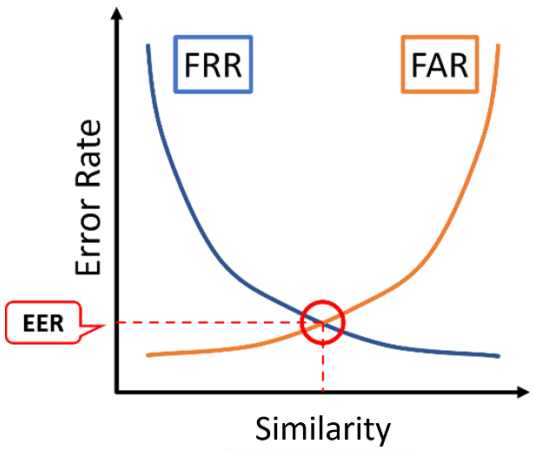


Figure 6. Ideal FRR-FAR curves image and definition of FRR and FAR in this study

4. Evaluations

In this section, three major tasks related to the evaluation of the proposed approach are discussed. First, how to establish the datasets for training and testing phases through the experiments is stated. Second, DP score threshold is practically determined in the training phase. Third, the performance of the proposed approach is evaluated by utilizing the testing dataset and the determined DP score threshold.

4.1. Data Acquisition

4.1.1 Experimental setup

To establish datasets for training and testing phases, the experiments were conducted. The datasets for each phase were actually human skeleton information extracted from monitored videos of photo-taking and net-surfing behaviors. As mentioned in **subsection 3.3**, OpenPose was utilized to analyze the monitored video in a fashion of frame-by-frame and extract the desired information of the human skeleton. In this study, we focused on three major parts ((I) upper arm, (II) lower arm, and (III) the angle of the bending arms) that can hypothetically differentiate the photo-taking behaviors from net-surfing behaviors. The experimental setup is described in Fig.7 where a user (assumed as a photographer) is taking action of either taking a photo or net-surfing using a smartphone, while the other is assumed as a photographed person. The whole behaviors of the photographer were continuously captured by another smartphone worn by the photographed person. In this experiment, the videos were taken from the right side of all participants as shown in Fig.7. Therefore, we hypothesized that there are remarkable differences in the movements of the participants' right arm. The videos were taken by Apple iPhone5s with a frame rate of 30 *fps*. 15 subjects were participating in this experiment to take action of either taking photos or net-surfing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

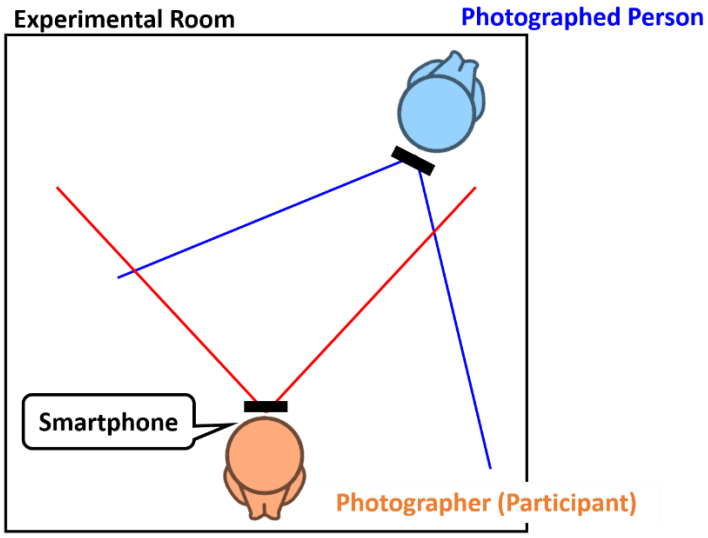


Figure 7. Experimental environment. The photographed person captures each behavior of the photographer (participant). The Photographer (participant) conducts photo-taking behaviors and net-surfing behaviors with a smartphone.

4.1.2 Experiment results

Fig.8 and Fig.9 partly illustrate the visual outputs obtained from OpenPose for photo-taking and net-surfing behaviors, respectively. More concretely, (a), (b), (c) and (d) in Fig.8 and Fig. 9 shows the first position frame before the behavior started, the frame when the behavior started, the frame during the behavior, and frame at the end of the behavior, respectively. The data obtained from the *xth* participant who performed a photo-taking behavior was denoted as "P*x*". Meanwhile, the data obtained from the *xth* participant who performed a net-surfing behavior was denoted as "N*x*". Thus, the dataset of photo-taking behaviors is ranged from P1 to P7, whereas, the dataset of net-surfing behaviors is ranged from N1 to N3. The details are tabulated in Table 3. Accordingly, in this paper, the dataset1, which consists of P1 to P7 and N1 to N3, was utilized for determining the DP score threshold. The dataset2, on the other hand, consists of P8 to P15 and N4 to N6 for testing the proposed approach.

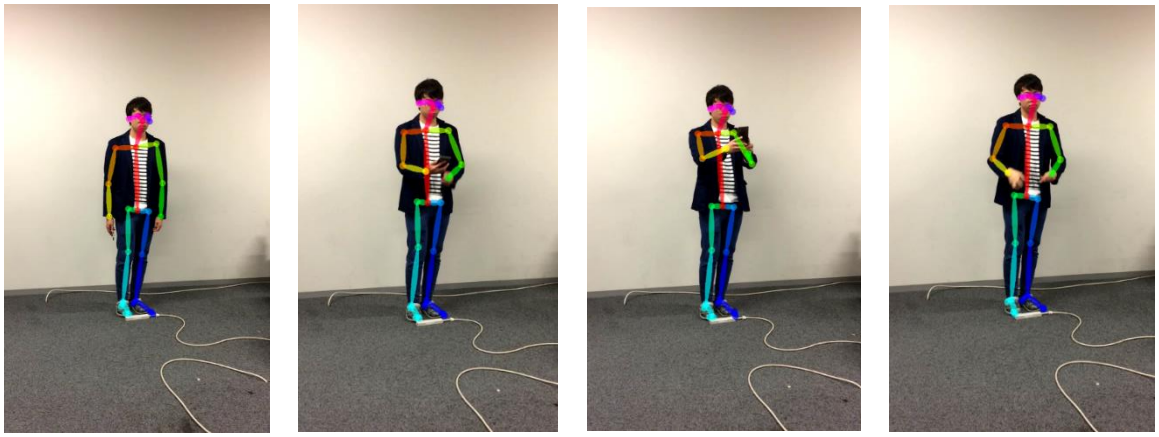
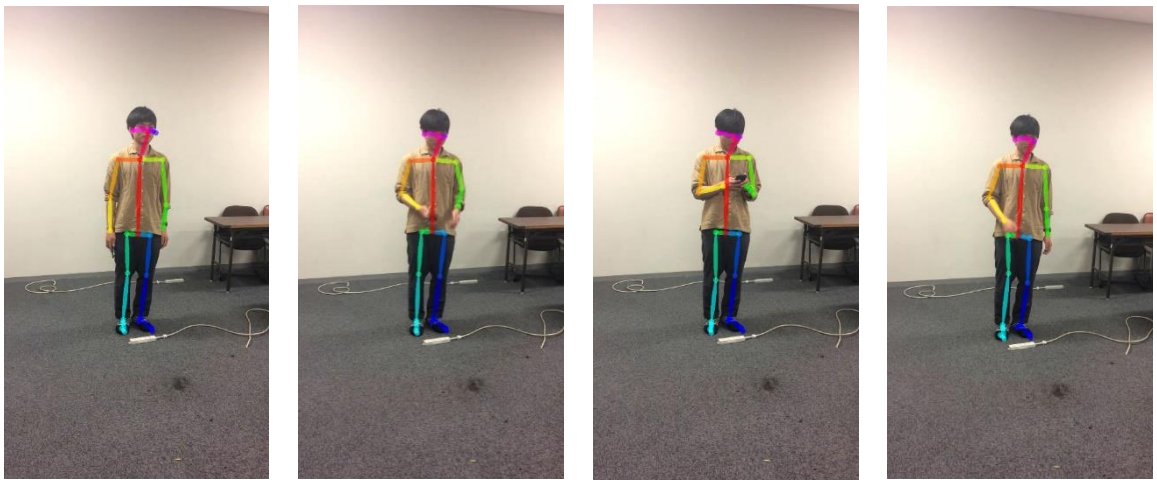


Figure 8. Result of output images from OpenPose for Photo-taking behaviors. (**a**) Image frame of the first position; (**b**) Image frame of when starting the photo-taking behaviors; (**c**) Image frame of when taking pictures by a smartphone (during photo-taking behaviors); (**d**) Image frame of when ending the photo-taking behaviors.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1



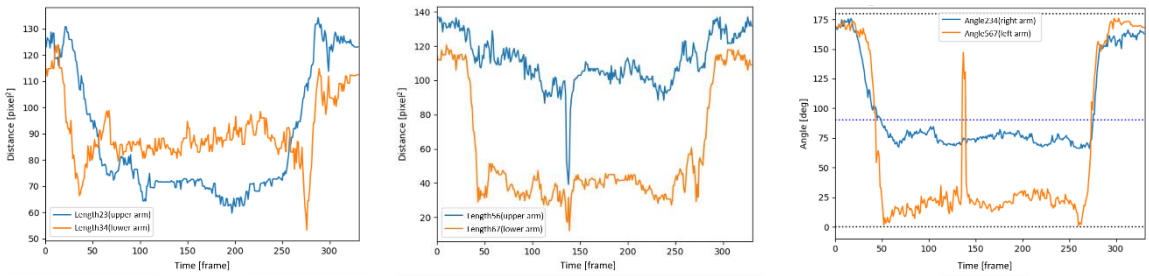
(a) (b) (c) (d)

Figure 9. Result of output images from OpenPose for Net-surfing behaviors. (**a**) Image frame of the first position; (**b**) Image frame of when starting the net-surfing behaviors; (**c**) Image frame of when surfing the Internet by a smartphone (during net-surfing behaviors); (**d**) Image frame of when ending the net-surfing behaviors.

Table 3. The obtained datasets from experiments

	Photo-taking behaviors	Net-surfing behaviors
Training phase (dataset1)	7 (P1, P2…, P7)	3 (N1, N2, N3)
Testing phase (dataset2)	8 (P8, P9…P15)	3 (N4, N5, N6)

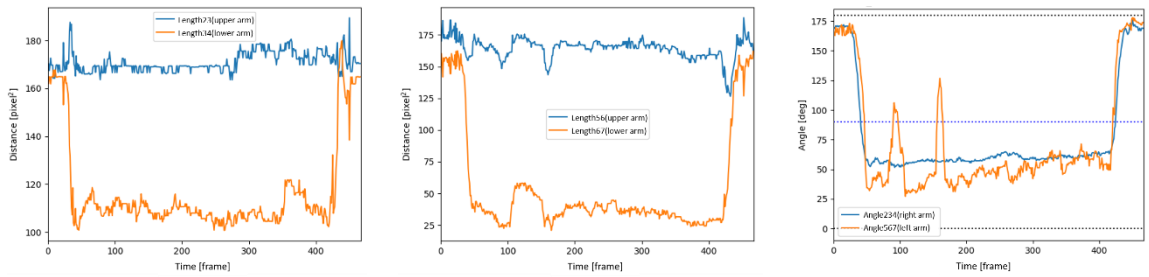
According to the output of OpenPose, Fig.10 and Fig.11 demonstrates the transitions of the considered three human parts (I, II, and III). While the transitions illustrated in Fig.10 were extracted from the OpenPose output data of P1 for a photo-taking behavior, those in Fig.11 were extracted from the one of N1 for a net-surfing behavior. In fact, for each subject, six joint components (joint 23, joint 34, joint 56, joint 67, angle 234, and angle 567) in total were considered for further analysis.



(a) (b) (c)

Figure 10. Transition results of arms length and angle for photo-taking behaviors from OpenPose (taken by P1). (**a**) upper arm lengths result of right and left arms; (**b**) lower arm length result of right and left arms; (**c**) angle results of bending right and left arms. In (a) and (b), vertical axis indicates distance between joints (length) [pixel²]. The horizontal axis indicates frame number it means time [frames]; In (c), the vertical axis indicates angle [degree]. The horizontal axis indicates frame number it means time [frames].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1



(**a**) (**b**) (**c**)

Figure 11. Transition results of arms length and angle for net-surfing behaviors from OpenPose (taken by N1). (**a**) upper arm lengths result of right and left arms; (**b**) lower arm length result of right and left arms; (**c**) angle results of bending right and left arms. In (a) and (b), vertical axis indicates distance between joints (length) [pixel²]. The horizontal axis indicates frame number it means time [frames]; In (c), the vertical axis indicates angle [degree]. The horizontal axis indicates frame number it means time [frames].

According to Fig.10 (b)(c) and Fig.11 (b)(c), there are no significant differences between the photo-taking and net-surfing behaviors in terms of the shape of the graph. However, the shape of the graph in Fig.10 (a) is quite different from the one in Fig.11 (a), which comes from the difference of the behaviors.

4.1.3. Data pre-processing

Even though the detection accuracy of human skeleton information using OpenPose[6] is very high, sometimes some of the joints are not correctly detected or even not detected, resulting in the need of pre-processing for the image frame. Figure 12 illustrates an example. Therefore, the numerical values of the joint components cannot be calculated. In such a situation, the coordinates of the undetected joint are predicted by performing interpolation using the frames before and after the considered frame.

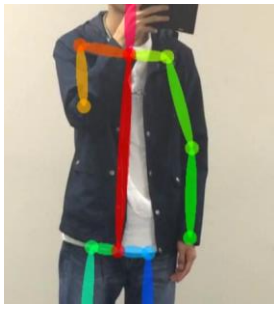


Figure 12. Example of non-detection by OpenPose

The misdetection, on the other hand, introduces a sudden change in the numerical values of investigating joint components as shown in Fig.10 (b) and (c). Figure 13 shows an example (Example 1) of misdetection result in the images and graphs. Fig.13 (a) and (b) are the same graphs as Fig.10 (b) and (c), respectively. In these graphs, the 139th frame in which a joint was mis-detected was emphasized by a yellow rectangle. Fig.13 (c) illustrates the images of the 135th, 139th, and 142nd frames.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

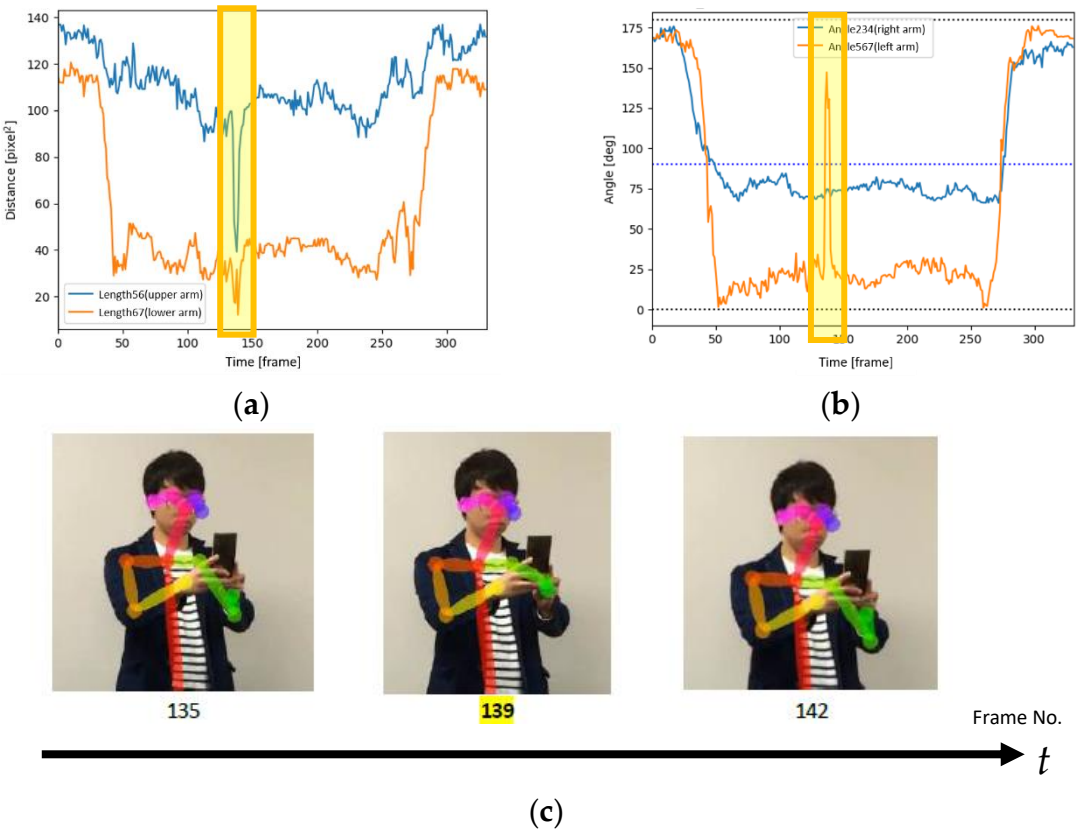


Figure 13. Example 1 of misdetection results image and graphs from OpenPose (taken by P1). (**a**) left arm's length results of the upper arm and lower arm; (**b**) angle results of bending right and left arms; (**c**) some results image in the focus part. In (c), the below number of Figures means the frame number of the video.

Fig.14 shows another example of misdetection (Example 2). In Fig.14 (a), the ideal estimated detection result is shown in white. As seen from the figure, the result obtained from OpenPose was different from the ideal estimated result. Figure 14 (b) shows some numerical values calculated from the coordinates of joints of the subject. The values fluctuate several times due to such a misdetection.

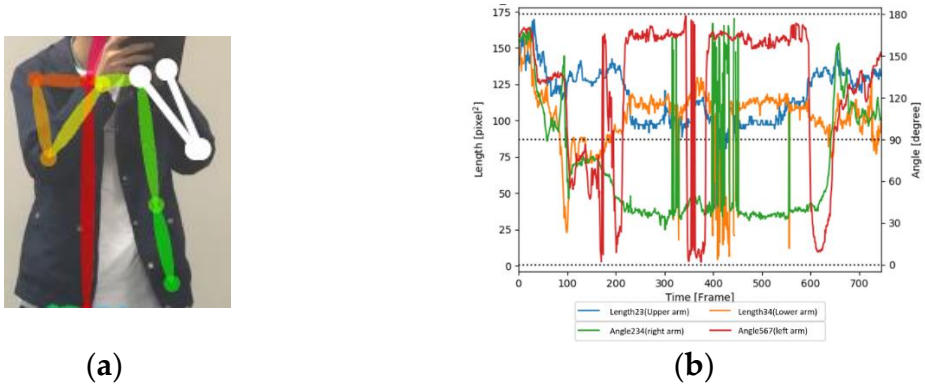


Figure 14. Example 2 of misdetection results image and graphs from OpenPose (taken by P5). (a) Output image frame (white line: ideal estimated detection result); (b) result of the right upper and lower arms lengths and the angle of bending right/left arm. In (b), first vertical axis indicates distance between joints (length) [pixel²]. Second vertical axis indicates angle [degree]. Horizontal axis indicates frame number corresponding to time [frames].

In fact, both non-detection and misdetection create noise in the obtained data. Before performing DP matching, it is necessary to smooth the data with as little noise as possible. Thus, in order to remove the noise, LPF (Low Pass Filter) was utilized. Thereby, we first extracted the frequency

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

components from the obtained data by using Fast Fourier Transform (FFT), then the cut-off frequency was assigned. In this case, a general Butterworth filter was used as the filter.

In this research, the desirable cut-off frequency of the LPF was determined as 40 [Hz] from the preliminary experiment. Figure 15 visualizes the processed data, which was done by LPF. This is actually the result of one of the participants in the photo-taking behavior. For the arm length, the vertical axis indicates the length [pixel^2], and the angle unit is shown in degrees. The horizontal axis indicates the frame corresponding to time. The line in red in each graph indicates the result of the filtering.

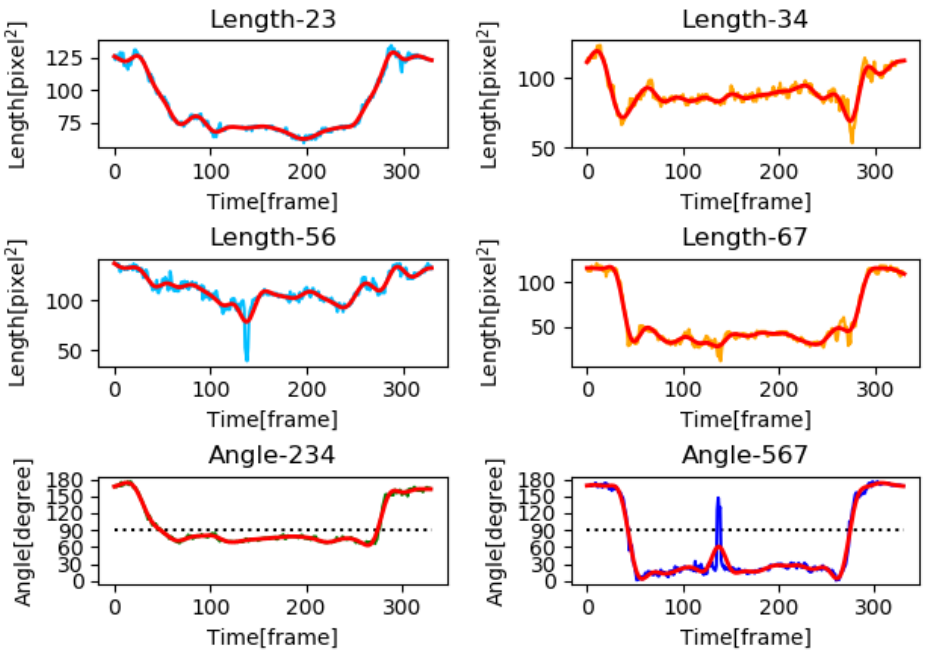


Figure 15. Result of each element for cutoff frequency 40Hz (taken by P1). Each element indicates the same as Table 2; Length-23 is right Upper arm; Length-34 is right Lower arm; Length-56 is left upper arm; Length-67 is left lower arm; Angle-234 is the angle of the bending right arm; Angle-567 is the angle of the bending left arm. Red line shows filtered value.

4.2. Determination of DP threshold

In order to determine the DP threshold, we first performed DP matching on the training dataset (dataset1). Since the data volume was small, the cross-validation approach was utilized. It means that the data of one of the participants was used as the referenced data in DP matching. Table 4 tabulates an example of DP matching output when data P1 was used as the referenced data. In this Table, a specific numerical value of DP score is assigned to each joint component. In fact, there were seven participants in our experiment, thus, seven tables of data which are the same type of table as Table 4 were obtained.

Table 4. DP matching results when P1 is used as the referenced data
(Reference data: P1, Input data: P2, P3, …, P7 and N1, …, N3)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

		length23	length34	length56	length67	Angle234	Angle567
Photo-taking Behaviors	P2	1.69	2.41	5.84	6.23	9.11	4.33
	P3	6.54	14.57	3.41	4.62	8.98	2.73
	P4	3.39	3.36	4.76	4.38	2.28	1.65
	P5	3.92	6.40	10.28	17.90	14.53	21.34
	P6	30.64	10.93	16.87	5.90	4.44	3.91
	P7	10.89	7.21	13.20	4.40	3.91	4.80
	Average	**9.51**	**7.48**	**9.06**	**7.24**	**7.21**	**6.46**
	S.T.	**9.89**	**4.20**	**4.84**	**4.82**	**4.15**	**6.74**
Net-surfing Behaviors	N1	58.88	8.50	28.60	2.04	3.48	6.31
	N2	21.48	3.18	11.17	3.26	2.73	8.20
	N3	29.05	3.27	9.50	1.25	1.39	6.94
	Average	**36.47**	**4.98**	**16.42**	**2.19**	**2.54**	**7.15**
	S.T.	**16.15**	**2.49**	**8.64**	**0.82**	**0.86**	**0.78**

As mentioned in **subsubsection 4.1.1**, the monitored videos were taken from the side of the photographer rather than from his/her front, thus, it is expected that not every joint component is equally important. Therefore, to select the important component for further investigation, the average DP scores with error bars of all joint components across all the participants are compared in Fig.16. Obviously, the right upper arm (length23) provides the most important feature of the average DP score, which shows the largest gap between the photo-taking and net-surfing behaviors. Since the higher the DP score is, the smaller the similarity between two data is, the result in Fig.16 indicates that only length23 among all the joint components is the potential joint for being considered in photo-taking behavior detection. Therefore, in this study, we focused on length23 not only for determining DP threshold but also for the performance evaluation of our proposed approach.

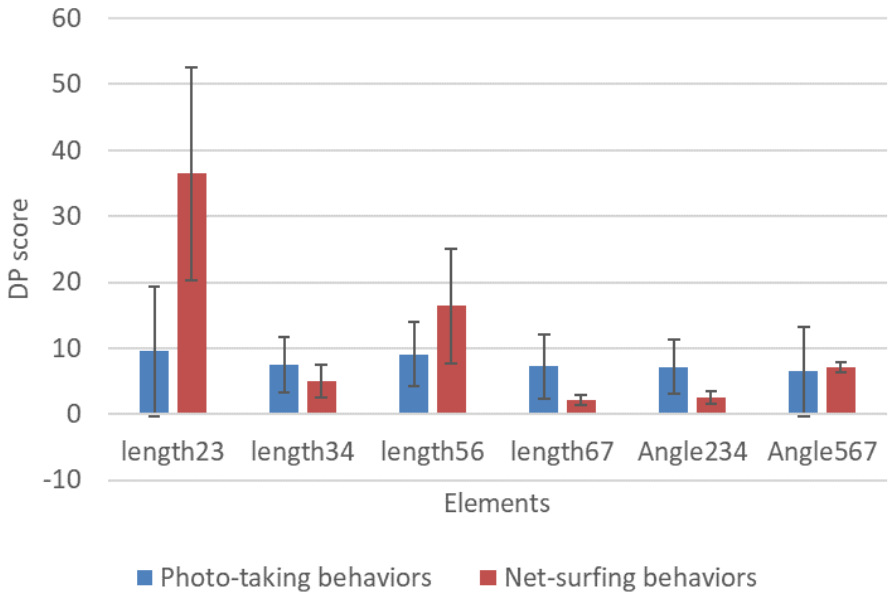


Figure 16. Average DP score for each behavior obtained from the results in Table 4 (Reference data: P1)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

By using DP score results for all photo-taking behaviors, the FRR and FAR can be calculated as follows:

FAR: Rate determined as photo-taking behaviors even though they are net-surfing behaviors in this study

$$\text{FAR} = \frac{\text{Number of mis-recognition as Photo-taking behaviors}}{\text{Number of all Net-surfing behaviors}} \quad (10)$$

FRR: Rate determined as net-surfing behaviors even though they are photo-taking behaviors in this study

$$\text{FRR} = 1 - \frac{\text{Number of correct recognition as Photo-taking behaviors}}{\text{Number of all Photo-taking behaviors}} \quad (11)$$

In this study, both FAR and FRR were plotted along with the so-called "assigned DP threshold" which was ranged from 0 to 35 with an increasing step of 2.5. If the DP score of each joint component in the referenced data of specific participant (as an example of P1 in Table 4) is less than "assigned DP threshold", it is determined that this participant performed a photo-taking behavior. Oppositely, a net-surfing behavior was determined when the DP score is more than "assigned DP threshold". Similarly, for seven participants who performed photo-taking behaviors, we could obtain seven graphs where both FAR and FRR were plotted. From the viewpoint of preventing appearance in photos as the research purpose, it is necessary to avoid non-detection (FRR) of photo-taking behaviors as much as possible. However, if the ratio of misdetection (FAR) increases with the priority given to that, the user must frequently deal with it when calling attention to the user, resulting in decreasing the usability. Figure 17 depicts two of seven graphs of the reference data P1 and P6. Thereby, seven values of EER were easily extracted from each graph. As mentioned earlier, EER is actually the intersection point of FAR and FRR where both of those values are equal. To determine DP threshold for our proposed approach, a general EER across all the participants must be obtained. For that reason, all seven EER values are plotted together in Fig.18.

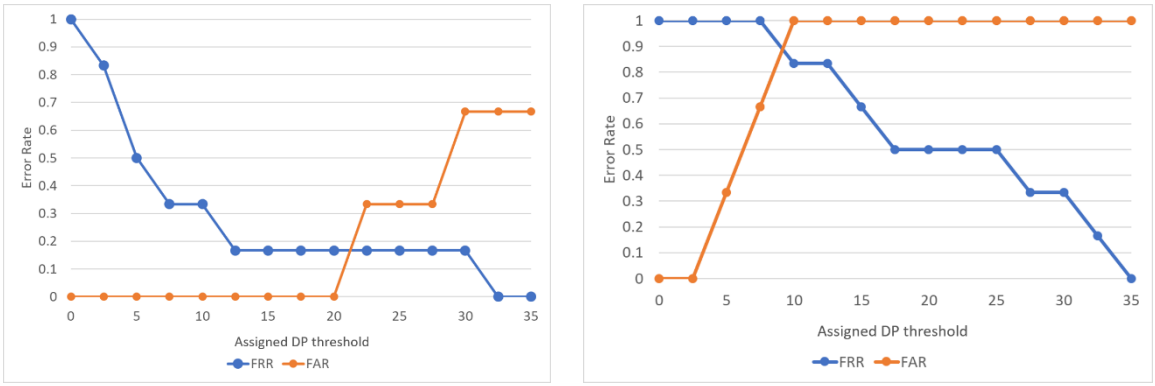


Figure 17. FRR-FAR curves. (**a**) Reference data: P1; (**b**) Reference data: P6. In each graph, the horizontal axis indicates assigned DP score threshold. The vertical axis indicates error rate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

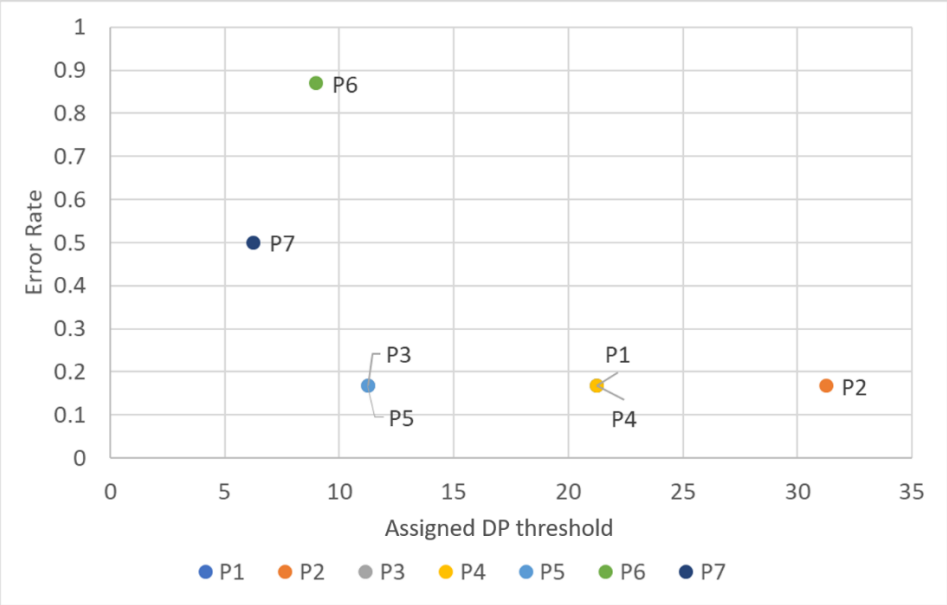


Figure 18. EER distribution obtained from all FRR-FAR curves by cross-validation for right upper arms (f_c=40Hz). The horizontal axis indicates the assigned DP score threshold. The vertical axis indicates the error rate. The legend indicates when using Photo-taking behaviors data as reference data.

According to Fig.18, most of EER values are less than 0.2 for photo-taking behaviors. On the other hand, the data with an obviously high error rate must be removed. Thereby, it is considered that two data with EER being higher than 0.4 are excluded. The average value of eligible EER was then calculated as about 0.17. Therefore, in accordance with the average value of 0.17 of EER, the DP threshold for our proposed approach was determined as about 15.9. In the next subsection, the performance of our approach is evaluated by using this DP threshold and dataset2.

4.3. Performance Evaluation of the proposed approach

In order to evaluate the proposed approach in the testing phase, DP matching was performed for the dataset2 using the dataset1 as the referenced data. The referenced data of outliers like P6 and P7 have been removed since the EERs are two high from Fig.19 as previously stated. The obtained DP scores were then compared with the DP threshold to identify photo-taking behaviors. The DP matching results are presented in Table 5. Note that, as explained in **subsection 4.2,** we only focused on the joint of length23, thus, Table 5 provides the results of DP matching with respect to this joint. In addition, the detection decision is expressed in cell colors. Accordingly, the yellow cells indicate that the behaviors were decided as photo-taking behaviors whose DP scores are less than the threshold (Th_{DP} = 15.9). In other words, the yellow cells show the correct decision using DP score. In addition, the light gray cells indicate the correct decision in net-surfing behaviors since those DP scores are more than the threshold. On the other hand, the gray cells with white numbers indicate the incorrect decision.

Table 5. DP matching result of dataset2 using dataset1 as the referenced data
(Reference data: dataset1 except outlier[P1, …, P5], Input data: dataset2[P8, …, P15 and N3, …, N6])

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2020 doi:10.20944/preprints202003.0376.v1

			Reference Data: Photo-taking behaviors (dataset1) length23(Upper right arm)				
			P1	P2	P3	P4	P5
Input Data (dataset2)	Photo-taking nehaviors	P8	8.10	7.89	5.69	8.27	3.40
		P9	4.60	6.68	3.84	4.04	5.91
		P10	0.93	2.84	5.17	2.96	3.28
		P11	8.44	9.49	2.07	3.33	3.22
		P12	1.14	2.30	5.74	2.81	2.91
		P13	5.61	6.00	1.42	1.93	3.65
		P14	21.10	28.43	7.01	17.08	10.18
		P15	13.59	13.74	10.18	13.51	3.14
	Net-surfing behaviors	N4	35.44	43.55	15.57	32.08	22.06
		N5	24.26	33.34	10.02	22.68	12.71
		N6	14.62	19.47	11.26	17.28	6.67

In this part, the detection accuracy of photo-taking behaviors should be focused on. According to Table 5, as a whole, the detection accuracy of photo-taking behaviors is 92.5%. Looking at the result in detail, when the reference data P1, P2 or P4 is used, the accuracy is 87.5%, while when the reference data P3 or P5 is used, the accuracy is 100%. On the other hand, the detection accuracy of net-surfing behaviors is 60%, which is not too high. However, the purpose of this study is to detect photo-taking behaviors with high accuracy in order to prevent unintended appearances in photos, and thus, the result of the performance evaluation is satisfactory.

In this proposal, OpenPose was used to obtain human's skeleton information and it performed effectively to provide each joint coordinate without any additional special equipment. In addition, Geometric feature analysis of human's joints elaborately worked to discriminate photo-taking behaviors and net-surfing behaviors with the support of DP matching. The DP threshold (Th_{DP}) was determined as 15.9 from the experiment, considering the characteristic of the DP score. Then, the Th_{DP} enabled to detect photo-taking behaviors with high accuracy which is more than 90%. This is because the focused human's body parts, namely, the lengths of upper arms changed differently between when performing photo-taking behaviors and when performing net-surfing behaviors. As a result, the proposed approach achieved very well.

5. Conclusions

In this paper, we proposed an approach which prevents unintended appearance in photos by analyzing human's behaviors in video data based on human's skeleton information obtained by OpenPose. The time-series data of the arm length and the angle of bending arm extracted from the human's joints information were focused on to detect photo-taking behaviors. The photo-taking behaviors are very similar to net-surfing behaviors in terms of smartphone's operation. Therefore, in this study, we tried to distinguish between these behaviors by cross-validation and DP matching. By focusing on the right upper arm length, the threshold of DP matching score, which is 15.9, was determined to detect the photo-taking behaviors.

Through the performance evaluation using the threshold of DP matching score, the proposed approach achieved the detection accuracy of 92.5%. Actually, if the obvious outlier data were not be used for calculating the accuracy, the detection accuracy would be 100%. Therefore, it is concluded that photo-taking behaviors can be distinguished from net-surfing behaviors with high accuracy, and the purpose of this study was sufficiently achieved by the proposed approach.

In this paper, detection of behaviors was mainly discussed. Therefore, the method to notify that someone is about to taking a photo must be discussed and clarify in future. Furthermore, it is important to implement the continuous DP matching in this approach to increase the detection accuracy as well as to achieve a real-time control for this research purpose.

References

1. We are social, “Digital in 2019”, https://wearesocial.com/global-digital-report-2019 (accessed on 31/01/2020).
2. Ministry of Internal Affairs and Communications (2018), “2018 WHITE PAPER Information and Communications in Japan,” Part2, Figure5-2-1-1, Transitions in household ownership rates for ICT devices, p.65.
3. Ministry of Internal Affairs and Communications (2018), “2018 WHITE PAPER Information and Communications in Japan,” Part1, Figure4-2-1-2, Terminals connected to the Internet, p.42.
4. Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven and Luc Vincent, “Large-scale privacy protection in Google Street View,” Proceedings on IEEE Computer Vision (ICCV2009), 2009, pp.2373-2380.
5. Takayuki Yamada, Seiichi Gohshi and Isao Echizen, “Evaluation of privacy visor to prevent face recognition from camera image (in Japanese),” The Institute of Electronics, Information and Communication Engineers Technical Report Order Information, Vol.112, No.420, EMM2012-92, 2013, pp.7-12.
6. Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," Proc. of Int’l Conf. on Computer Vision and Pattern Recognition, pp.1302-1310, 2017.
7. Tatsuya Koyama, Yuta Nakashima and Noboru Babaguchi, “Real-time privacy protection system for social videos using intentionally-captured persons detection,” Proceedings IEEE of Multimedia and Expo (ICME2013), 6 pages, 2013.
8. Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, June 2016, pp.779-788.
9. Yuhi Kaihoko, Manami Kanamaru, Yoshihiro Nagano, Phan Xuan Tan and Eiji Kamioka, “Prevention of Unintended Appearance in Photos through Detection of Taking Photo Behaviors by Photographer (in Japanese),” The Institute of Electronics, Information and Communication Engineers Technical Report Order Information, Vol.118, No.305, MoNA2018-34, November 16, 2018, pp.69-74.
10. Hsin-Chun Tsai, Chi-Hung Chuang, Shin-Pang Tseng, Jhing-Fa Wang, "The Optical Flow-based analysis of human behavior-specific system," 2013 1st International Conference on Orange Technologies (ICOT), 2013, pp.214-218.
11. Yuhi Kaihoko, Tan Phan Xuan, Eiji Kamioka, "Identification of Photo-taking behaviors using Optical Flow Vector," International Journal of Advanced Trend in Computer Science and Engineering, Vol. 8, No. 1.4, 2019, pp.306-312.