17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Position Estimation of Pedestrians in Surveillance Video Using Face Detection and Simple Camera Calibration

Toshio Sato, Xin Qi, Keping Yu, Zheng Wen, Yutaka Katsuyama
Global Information and Telecommunication Institute,
Waseda University,
Tokyo, Japan
toshio4.sato@aoni.waseda.jp

Takuro Sato
Research Institute for Science and Engineering, Waseda University,
Tokyo, Japan
t-sato@waseda.jp

Abstract

Pedestrian position estimation in videos is an important technique for enhancing surveillance system applications. Although many studies estimate pedestrian positions by using human body detection, its usage is limited when the entire body expands outside of the field of view. Camera calibration is also important for realizing accurate position estimation. Most surveillance cameras are not adjusted, and it is necessary to establish a method for easy camera calibration after installation. In this paper, we propose an estimation method for pedestrian positions using face detection and anthropometric properties such as statistical face lengths. We also investigate a simple method for camera calibration that is suitable for actual uses. We evaluate the position estimation accuracy by using indoor surveillance videos.

1. Introduction

To enhance video surveillance system functions, pedestrian position estimation is expected for many applications, such as tracking persons [1], indoor positioning where the GNSS does not work [2], making triggers such as proximity sensors [3], controlling wireless communication [4], and extracting partial images for effective communication [5]. Although other method such as LIDAR can be used to measure the position, the application will be expanded if the position estimation can be realized using only video images at the location where the camera is installed. Detection and segmentation of the human body are widely used for position estimation and camera calibration [1][6][7][8]. In the case of close-up camera settings, the whole body may sometimes protrude from the field of view, and body detection cannot be applied. It is obvious that face detection has higher detection opportunities in surveillance videos.

Accuracy is important for position estimation. Reference [2] reported that positioning errors are 1 m to 2 m by using a direct linear transform. Reference [1] reports that the average error of position estimation for distance becomes 64 mm with the variation of 414 mm by using camera autocalibration. Although precise camera calibration is important and there are many studies on camera autocalibration using pedestrians captured in video data, experimental verification of position estimation is limited.

Simplicity of camera calibration is also important. Reference [1] uses camera autocalibration to calculate parameters except for the focal length. The focal length is always difficult to determine if a varifocal lens is used. Reference [5] used another approach that calculates a perspective projection matrix without using a camera model; however, four points on the floor should be marked to know exact positions.

In this paper, we investigate a practical method for pedestrians position estimation in surveillance videos in a corridor. Setting a video camera in the corridor often makes it easier to capture faces. Face detection with anthropometric properties such as a statistical face length is used to estimate pedestrian foot positions. This approach does not use body detection and can estimate the position at an area where the whole body is not captured in the frame. Furthermore, simple camera calibration is discussed without using prior knowledge of the focal length and height of the camera for already existing surveillance cameras.

2. Methods

We assume that a surveillance camera is placed in a typical installation, as shown in Figure 1(a). The camera is placed at the height yc and the tilt angle Θ. The world coordinate (x,y,z) is defined with the origin (0,0,0) at the floor position of the camera. We define the pedestrian height as yi at distance z_i. In this study, the roll angle is assumed to be zero, and lens distortion is ignored.

Pedestrians on the world coordinate system are projected to a camera image plane that is represented by coordinates (u,v), as shown in Figure 2(a). One pedestrian is observed from the head position v_h to the foot position v_b for the v-axis. Although we use a video camera that captures a 3264x2448 image, a virtual larger image is used to handle the left person's case in Figure 2(a). In Figure 2(a), v_0 and v_c are coordinate values for the vanishing line and the center of a captured image, respectively.

The relationship between the world coordinate system (x,y,z) and the image coordinate (u,v) is represented by the perspective projection of Equation 1 [1].

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z} \begin{bmatrix} f & 0 & u_c \\ 0 & f & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & y_c \\ 0 & \sin\theta & \cos\theta & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (1)$$

In this paper, we focus on (y,z) in the world coordinates. Using the second equation in Equation 1 with

conditions of $v = v_b$ and $y = 0$, the position z_i at the foot position (y=0) is calculated as Equation 2[1].

$$z_i = \frac{f\, y_c}{f \sin\theta - (v_b - v_c)\cos\theta} \qquad (2)$$

The tilt angle θ is calculated in the image coordinates by using the focal length f, as shown in Figure 1(b).

$$\theta = \tan^{-1}\left(\frac{v_c - v_0}{f}\right) \qquad (3)$$

While v_c corresponds to the center of the captured image (v_c=1224), other parameters v_0, y_c and f are unknown in Equations 2 and 3. The foot positions v_b should be estimated by using the face detection results.

Figure 3 illustrates the flow of our proposed method to estimate parameters and the z position of pedestrians. To detect foot positions from the face detection results, we use anthropometric properties that are the statistical length of bodies. The camera height y_c and the vanishing line v_0 are estimated by using multiple detection results for pedestrians in the video. The focal length f is calculated by using several frames that capture a person at an arbitrary distance. Finally, position estimation is executed for pedestrians in surveillance videos. Detailed descriptions follow from the next section.

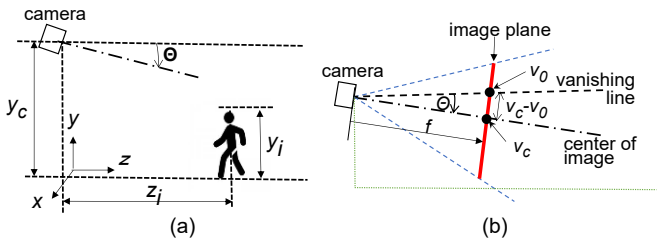

Figure 1. (a) Arrangement of a surveillance camera in the world coordinate system. (b) image plane.

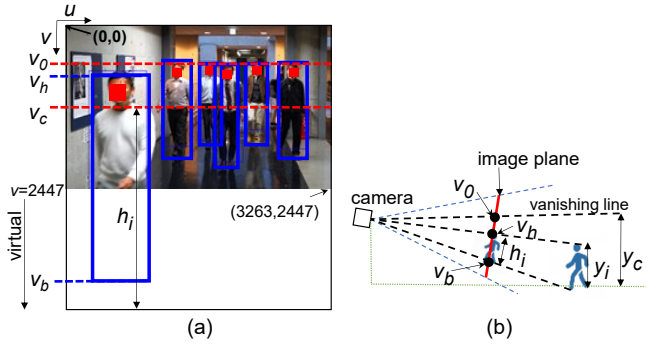

Figure 2. Image coordinate (u,v) and parameters.

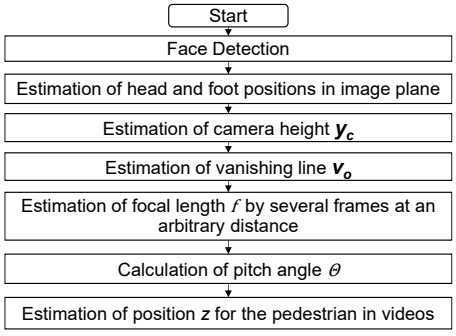

Figure 3. Process flow of the proposed method

2.1. Face Detection

Instead of using body detection, we employ the Viola-Jones face detection method [9] to estimate the foot position. One face detection result includes information about an origin point (u_f,v_f) and size (L,L), as shown in Figure 4(a).

2.2. Position Estimation in the Image Plane Using Anthropometric Properties

The pedestrian's foot position (u_b,v_b) in an image is estimated by using face detection results and anthropometric properties. As shown in Figure 4(a), the vertical coordinate value v_f and the height L of the face detection rectangle are used to calculate the foot position v_b by

$$v_b = v_f + k \cdot L \qquad (4),$$

where k is the proportion coefficient. Based on anthropometric properties for Japanese individuals [10], the average height of adolescents is 1,699.1 mm, and the average morphological face height is 121.1 mm, as shown in Table 1. The morphologic face height is the distance from the eyes to the lowest point of the chin, as shown in Figure 4(b). We consider that it is almost the same length of the face detection result from eyebrows to the bottom of the mouth. According to the relationship of human dimensions in Table 1, k is simply considered 13.3 (=(1699.1–91.0)/121.1), where 91.0 is the distance from the glabella (the position between eyebrows) to the vertex (the crown), as shown in Figure 4 (b) and Table 1.

The head position in an image v_h is also calculated using the same approach as v_b. The parameter m of Equation 5 is obtained as m=0.75 using the values in Table 1.

$$v_h = v_f - m \cdot L \qquad (5)$$

The horizontal foot position u_b is taken to be the midpoint of the face rectangle.

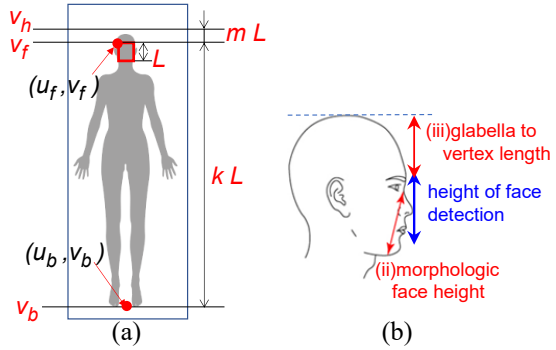

Figure 4. Estimation of position in an image. (a) Relationship between face detection and foot position. (b) Face length [10].

Table 1. Anthropometric properties [10].

	Number of samples	Mean (mm)	Standard deviation (mm)
(i) body height	56	1699.1	58.6
(ii) morphologic face height	56	121.1	6.4
(iii) glabella to vertex length	47	91.0	6.9

2.3. Camera Calibration

The vanishing line corresponds to the projection of a point at infinity where the pedestrian's height becomes zero. Figure 5 shows the relationship between the estimated foot position v_b and the height (v_b-v_h) in the image coordinates. In Figure 5, foot positions over 2,447 are included that are estimated in the virtual image plane. The vanishing line v_0 is obtained by approximating the plot data with a linear equation and calculating the intersection point of the height as zero. According to the results in Figure 5 that are measured by using the six persons walking video, the vanishing line v_0 is estimated as 784.6 pixels from the top of the image.

The camera height is obtained by the ratio between the camera height and pedestrian height. In the image coordinates, the camera height projection is assumed to be v_0, and the relationship between the world coordinates is described as

$$\frac{y_c}{y_i} = \frac{v_b - v_0}{h_i} \quad (6),$$

where y_i is given by the body height in Table 1 and h_i is the height in an image (see Figure 2(a) and (b)).

Figure 6 shows the relationship between foot positions v_b and y_c calculated by Equation 6. The total camera height y_c is estimated as 2,172.3 mm, approximating the plot data with a linear equation and calculating the intersection point of $v_b = v_0$.

The focal length f is determined by analyzing images including a person standing at an arbitrary distance. This arbitrary fixed distance z_a is only one parameter that we need to measure in the proposed method. The resolution for distance z_a/f (mm/pixel) is equal to the resolution of the pedestrian's height, as shown below.

$$\frac{z_a}{f} = \frac{y_i}{v_b - v_h}$$

Although z_a can be selected arbitrarily, in this paper, we acquire 48 frames at z_a = 6,000 mm (see the example of the image in Figure 7(e)) and calculate the average of (v_b-v_h) as 1,453.8 pixels. The focal length is estimated as 5,130 pixels.

Table 2 illustrates the given parameters, including z_a, to calculate f. Other given parameters are obtained from the dimensions of captured images and anthropometric properties. Table 3 indicates the estimated parameters by using our method. They are always difficult to measure precisely in practical installation.

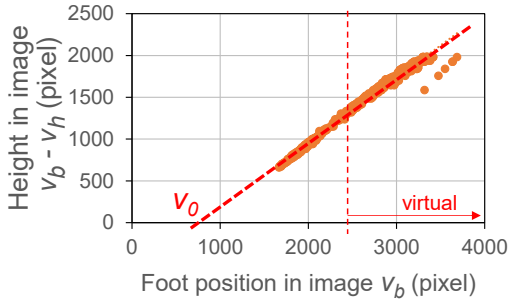

Figure 5. Estimation of vanishing line v_0.

Table 2. Parameters to be given.

Parameters	Given values
The center of the image (v_c)	1,224 pixel (the center of a captured image)
Distance to calculate focal length (z_a)	6000 mm
Body height (y_i)	1,699.1 mm (Table 1)

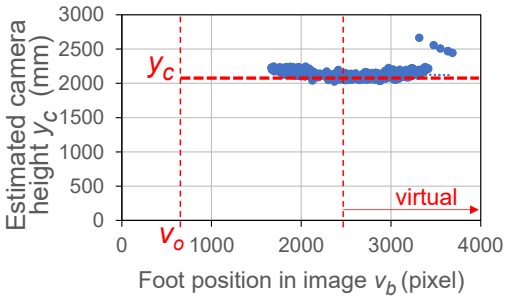

Figure 6. Estimation of camera height y_c.

Table 3. Estimated results for parameters.

Parameters	Estimated values	Known values
Vanishing line (v_0)	784.6 pixel	N/A
Camera height (y_c)	2,142.0 mm	2 m + α . [approx. 2 m (tripod stand) + attachment + 19 mm (camera device)]
Focal length (f)	5,130 pixel (for 2448 pixel of image height)	2,147-21,474 pixel. [using a varifocal lens of 5 – 50mm]

3. Experimental Results

We evaluate our proposed method using a 92 s video (resolution of 3264x2448 and frame rate of 15 fps) in which six pedestrians walk indoors. A target person stops every 1 m at distance z, as shown in Figure 7. In Figure 7(f), the foot position cannot be observed in the captured image. The target person determines the distance confirming scale marks on the floor. For this target person, 160 frames are detected through movement.

The developed process detects face regions and calculates the position z for each frame based on Equations 2 and 3 with parameters in Tables 1 through 3. The parameters in Table 3 are calculated by using whole frames in video data that include six pedestrians. Figure 8 illustrates the relationship between the measured (ground truth) position and estimated results. The estimated position tends to decrease for larger distances over 12 m. As shown in Figure 7(f), position estimation is possible even when the whole body does not appear in the image. Table 4 shows the root mean square errors (RMSEs) of the estimated positions for the 12 positions. In Table 4, the results of a comparative method that also uses perspective transformation without camera calibration by determining four positions on the floor [5] are included. The perspective transformation matrix A in an equation

$$\begin{bmatrix} x \\ z \\ 1 \end{bmatrix} = A \begin{bmatrix} u_b \\ v_b \\ 1 \end{bmatrix}$$

is calculated by four points (x,z) on the floor and (u,v) in the image plane, as shown in Figure 9.

Table 4. Estimation errors at 12 positions.

Grand Truth z (mm)	4000	5000	6000	7000	8000	9000	1000	11000	12000	13000	14000	15000	Average
Number of face detection	1	11	48	21	17	35	18	14	1	7	4	1	
RMSE of 4-point perspective (mm)	670.0	1202.5	898.5	793.9	423.9	559.8	427.9	510.5	240.0	1401.0	1808.1	2020.0	822.0
RMSE of proposed method (mm)	109.9	424.3	323.3	258.4	427.1	515.1	394.2	504.8	113.9	1040.6	1488.6	1766.8	512.4

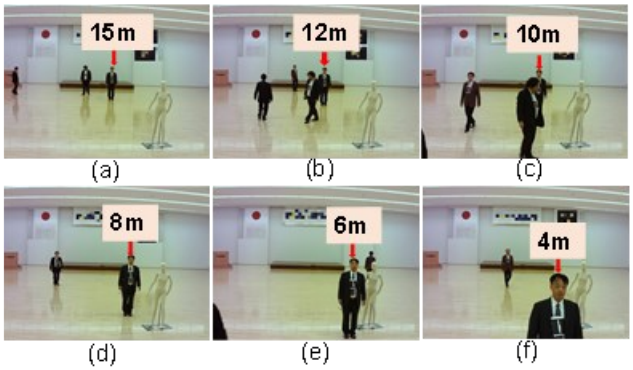

Figure 7. Examples of the video for evaluation. Some regions are masked for privacy.

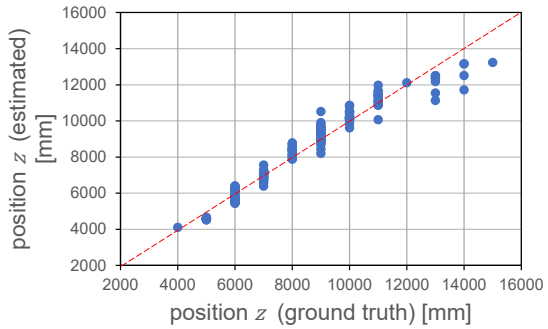

Figure 8. Estimated z for 12 positions

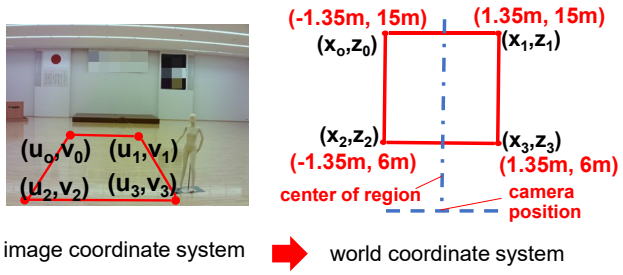

Figure 9. A method to calculate a matrix for four-points perspective (a method for comparison).

According to the results in Table 4, the average RMSE of all data is 512.4 mm, which is superior to the 822.0 mm of the four-point perspective approach. While the four-point perspective has larger errors not only over 12 m but also under 7 m of distance, the proposed method has smaller errors under 7 m. The results indicate that our approach can estimate the position of pedestrians with equivalent precision by using a simple calibration method.

4. Discussions

We propose a method to estimate pedestrians' positions by using face detection instead of body detection. By estimating the head and foot positions in a virtual image plane, it is possible to determine position z on the floor even when it extends below the screen, as shown in Figure 7(f), and another person's occlusion occurs, as shown in Figure 7(c). In this experiment, there was no face detection at position z of 3 m and only one frame of face detection at the 4 m position. We can expand the range of position estimation by applying advanced face detection [11].

According to the results of Table 4, the error at the position of 12 m is small; however, this is not certain due to the small number of face detections. Figure 8 shows that our approach shows sufficient accuracy of estimation over a wider range at a distance of 11 m. The size of the detected face area L was less than 50 pixels at a distance of more than 12 m. This might be one of the reasons for larger errors over 13 m. If we can use higher resolution devices, these errors may be decreased.

The accuracy of face detection appears as variation in the same position z in Figure 8. The variation is beyond 2.5 m (30%) at z=9 m. By using the average of multiple frames, these errors simply improve. Moreover, the variation in anthropometric properties should be considered. According to Table 1, the standard deviation scores are 3% to 8% for the measurement values. Although reference [5] reports that six Asian persons have similar position estimation accuracy, a much wider diversity of people should be discussed in our future work. Overall, the accuracy of 0.5 m (5% errors at the position of 11 m) is regarded to be close to the limit of this approach.

We adopt simple calibration for already installed surveillance cameras to estimate pedestrian positions. Our method uses several frames that include a pedestrian standing at a specific distance to calculate the focal length. Other parameters are obtained by analyzing pedestrians in the surveillance video. Although we focus on the estimation of the position z, the estimation of the position x is also possible by solving Equation 1.

There are many studies on camera autocalibration using pedestrians [6][7][8]. By applying these studies, it is expected that all parameters can be automatically determined and the pedestrian's position can be estimated.

In this study, we focus on surveillance videos in a corridor that make it easier to capture faces. In the case of free walking, face detection may fail. It may be effective to use our proposed method and the body detection method together to realize robustness.

5. Conclusion

We proposed a simple pedestrian position estimation in surveillance videos using face detection and anthropometric properties. Camera calibration easily obtains parameters for already installed surveillance cameras by capturing several frames at an arbitrarily distance. We confirmed that the estimation errors are almost less than

0.5 m at positions from 4 m to 11 m through the experiment. The result indicates that our approach is a simple method to estimate the position of pedestrians that is applicable to existing surveillance cameras. We will continue to pursue fully automatic camera calibration to estimate the position of pedestrians.

Acknowledgments

This work was supported by a research grant for expanding radio wave resources (JPJ000254) of the Ministry of Internal Affairs and Communications under the contract for “Research and development of radar fundamental technology for advanced recognition of moving objects for security enhancement”.

References

[1] H. Ando and H. Fujiyoshi: “A Method for Estimation of 3D Position and Camera Self-calibration Using Results of Human Detection,” The Transactions of the Institute of Electrical Engineers of Japan. D, vol.131, no.4, p.9 & pp.482-489, 2011.

[2] T. Tsai, C. Chang, S. Chen, and C. Yao: “Design of Vision-based Indoor Positioning based on Embedded System,” IET Image Processing, vol. 14, no. 3, pp. 423-430, 2020.

[3] K. Yu, X. Qi, T. Sato, S. H. Myint, Z. Wen, Y. Katsuyama, K. Tokuda, W. Kameyama, and T. Sato, “Design and Performance Evaluation of an AI-based W-band Suspicious Object Detection System for Moving Persons in the IoT Paradigm,” IEEE Access, vol. 8, pp. 81378–81393, 2020.

[4] M. Alrabeiah, A. Hredzak and A. Alkhateeb: “Millimeter Wave Base Stations with Cameras: Vision-Aided Beam and Blockage Prediction, ” 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, pp. 1-5, 2020.

[5] T. Sato, X. Qi, K. Yu, Z. Wen, S. H. Myint, Y. Katsuyama, K. Tokuda, T. Sato: “Pedestrian Positioning in Surveillance Video using Anthropometric Properties for Effective Communication, ” 23rd International Symposium on Wireless Personal Multimedia Communications (WPMC), Okayama, Japan, pp. 1-6, 2020.

[6] F. Lv, T. Zhao and R. Nevatia; “Camera Calibration From Video of a Walking Human, ” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 9, pp. 1513-1518, 2006.

[7] I. Junejo and H. Foroosh: “Robust Auto-Calibration from Pedestrians, ” 2006 IEEE International Conference on Video and Signal Based Surveillance, Sydney, NSW, Australia, pp. 92-92, 2006.

[8] W. Kusakunniran, H. Li and J. Zhang: “A Direct Method to Self-Calibrate a Surveillance Camera by Observing a Walking Pedestrian,” 2009 Digital Image Computing: Techniques and Applications, Melbourne, VIC, Australia, pp. 250-255, 2009.

[9] P. Viola and M. Jones: ‘‘Rapid Object Detection Using a Boosted Cascade of Simple Features,’’ IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, p. I-I, 2001.

[10] M. Kouchi and M. Mochimaru: ‘‘AIST Anthropometric Database,” National Institute of Advanced Industrial Science and Technology,’’ H16PRO 287, 2005. <https://www.airc.aist.go.jp/dhrt/91-92/index.html>

[11] D. Triantafyllidou and A. Tefas: ‘‘Face Detection Based on Deep Convolutional Neural Networks Exploiting Incremental Facial Part Learning,’’ 23rd International Conference on Pattern Recognition (ICPR), pp. 3560-3565, 2016.