Person Re-identification Based on Body Segmentation

Hua Jiang and Liang Zhang(✉)

Tianjin Key Lab of Advanced Signal Processing,
Civil Aviation University of China, Tianjin, China
l-zhang@cauc.edu.cn

Abstract. Person re-identification is a difficult problem to solve in the process of video analysis of non-overlapping multi-camera surveillance system. A new algorithm of person re-identification is proposed in the base of the human segmentation parts in this article. First, the human body segmentation is achieved based on the depth of bone points. Second, the optimal key frame is selected by using the scoring strategy for all parts of the same human multi-frame images segmentation. Different weights for the global color feature and the HOG feature is assigned. Third, all the characteristics are combined to establish a human target model, and EMD (Earth Movers Distance) distance is used to determine the similarity between the targets. The Kinect REID and BIWI RGBD-ID databases are used in the experiments. The results show that the proposed method has stronger robustness and a higher recognition rate.

Keywords: Person re-identification · Human division
Depth information · Color characteristics · HOG characteristics

1 Introduction

The human body re-recognition is the basic research work of pedestrian gesture, action and behavior recognition which goal is to correlate pedestrian images obtained from multiple cameras, and then judge these from the different people who are estimated images of the human body are the same person [1]. Due to its non-contact characteristics, this technique has broad application prospects in monitoring video data processing, automatic photo annotation and image retrieval.

Early re-identification studies focused on two directions: one is based on the feature representation of the method. The other is based on the distance measurement method [2]. The feature class method aims at designing strong differentiation and stability features. This method improves the accuracy of re-recognition to a certain extent, but it only considers the human target from the overall characteristics, and lacks the spatial binding information about the human target. The measurement method requires a low design requirement for the feature, from the perspective of the measurement distance. Its performance

J. Yang et al. (Eds.): CCCV 2017, Part II, CCIS 772, pp. 474–485, 2017.
https://doi.org/10.1007/978-981-10-7302-1_39

largely depends on the selection of the sample. When sufficient samples are available, the distance function learned can be generally applied to the recognition problem in a variety of environments. And when the number of samples are small, there will be a fitting phenomenon. In addition, the training data samples need to manually annotate and consume a lot of labor and time costs [3].

In the traditional pedestrian re-recognition methods, pedestrian appearance is usually viewed as a whole for modeling and matching. In order to improve the accuracy of human re-recognition, in the identification of the use of local matching based on the use of pedestrian global appearance processing architecture. The method can better cope with the impact of pedestrian appearance of the deformation. To this end, the researchers proposed a variety of block based pedestrian re-recognition method. Alahi [4] et al. Used a set of rectangles range large from small to divide the human body into multiple rectangular regions, and then combined each local feature to construct a new appearance model. Marimon [5] et al. extracted the characteristics of the overall image of the pedestrian, and then divided it into four equal parts, nine equal points, sixteen equal parts, and then calculate the characteristics. Bak [6] et al. Maximized the distance between the major color sets of the upper and lower parts of the body, dividing the body into the upper and lower parts, and using the body position detector to find a meaningful sub-region. SDALF (Symmetry-Driven Accumulation of Local Features) method [1]. First, the human body is divided into three parts of the head, trunk and legs, and then use the body's symmetry were the trunk and legs which were divided, according to the symmetry axis extraction weighted HSV color histogram, the MSCR (Maximified Stable Color Regions) feature, and the RHSP (Recurrent High-Structured Patches) feature. But this division does not guarantee accurate part color. For example, if the person's shirt is longer and into the leg portion, then the color characteristics of the leg will be contaminated by unwanted shirt colors.

This paper presents a method of human re-identification based on body segmentation. First, the human body segmentation is performed that it is based on the depth of bone points. The optimal frame is selected for all parts of the multi-frame images using the position scoring strategy. And then fused the global color histogram and HOG characteristics of the body parts of the characterization. The experimental results show that the pedestrian re-recognition method based on block matching has higher robustness and higher recognition precision.

2 Partial Segmentation and Optimal Frame Selection

Body segmentation means that the human image is divided into several meaningful parts according to the human body structure. Proper and accurate segmentation not only can improve the recognition accuracy, but also reduce the complexity of the subsequent image processing. In this paper, the segmentation process is shown in Fig. 1.

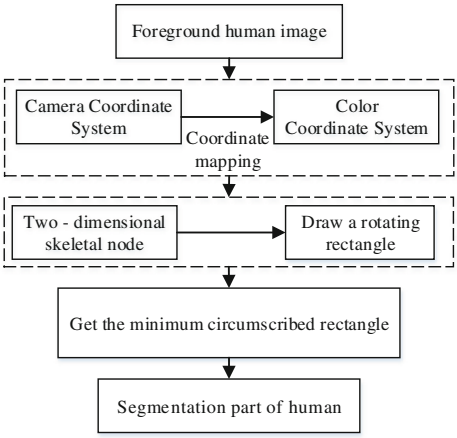

Fig. 1. Flow chart of human division

2.1 Segmentation Based on Depth of the Skeletal Point

For the foreground image that has been denoised, the segmentation of this paper is based on the position information of 20 skeletal nodes corresponding to the human body in the Kinect Camera coordinate system [7], as shown in Fig. 2.

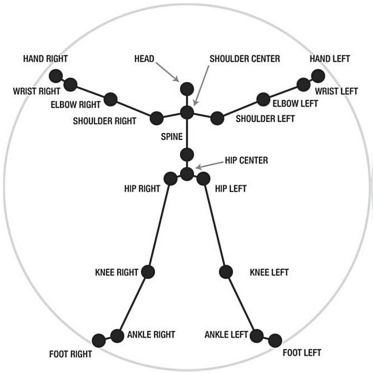

Fig. 2. Marked with 20 skeletal nodes of the human body

2.1.1 Draw a Rotating Rectangle

Select the location of the two ends of the skeleton nodes, the distance between the nodes are long of the rectangle. According to the experiment and experience. In addition to the width of the trunk part of the long two-thirds, other parts of the width are set to a long one-half. The midpoint of the coordinates of the two skeletal nodes are set as the rotation center of the rotating rectangle, and the

rotation rectangle is drawn with the angle between the connection of the two skeletal nodes and the horizontal plane, as shown in Fig. 3.

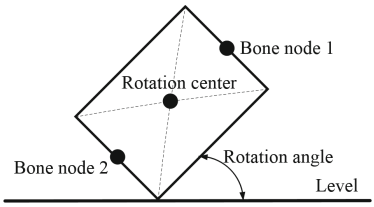

Fig. 3. Draw rotating rectangle

2.1.2 Partial Segmentation

Modern clothes have many styles and colors, the color of the sleeves and the torso is often part of the pattern of color which is not the same, the pants of the thigh and calf at the color may not be the same. And the head area is relatively subtle, not easy to extract effective information, feet area is easily blocked. Based on the above considerations, except to the head and the foot. The whole body is divided into 9 parts which are trunk, right upper arm, right lower arm, left upper arm, left lower arm, right thigh, right calf, left thigh, and left calf. There are 9 pictures in the human body on the rectangle inside the box is part of the split out of the human body, as shown in Fig. 4.

Fig. 4. 9 Parts after the human body segmentation

2.2 Optimal Frame Selection Based on Site Scoring

Because of the human always moving in the shooting process. The angle and gesture of the Kinect relative to the moving target are different greatly during the period of time. Different parts may be blocked at different times.

So the integrated multi-frame image, the block or other reasons for the loss of the use of other parts of the frame image to complement the corresponding parts. The use of site scoring strategy, the body of the 9 parts was selected optimal image preservation.

Among them, the site score using the following three indicators

(1) The ratio of the non-background area to the entire pair of images
Traverse the entire image, the pixels are as white as the background part, the sum is represented by Δ, The foreground is indicated by sum_2, set the foreground part of the ratio of the whole image is Δ, which is

$$\Delta = \frac{sum_2}{sum_1 + sum_2} \tag{1}$$

(2) Angle difference of human body
The human target in the Kinect of Camera coordinate system is shown in Fig. 5.
The x, y, z direction of the angle difference were recorded as α, β, γ that set the total offset of χ, in order to illustrate the effectiveness of the algorithm, the two images of the same human body in different sets are selected under different illumination and angle conditions, and two different images of the same human body have a viewing angle of 90 to 180°, So the query set and the candidate set of the human body image selection have the biggest difference in the shooting angle, which is

$$\chi = \begin{cases} \frac{180-\alpha}{180} + \frac{180-\beta}{180} + \frac{180-\gamma}{180}, & \textit{Query set} \\ \frac{\alpha}{180} + \frac{\beta}{180} + \frac{\gamma}{180}, & \textit{Candidate set} \end{cases} \tag{2}$$

(3) Image quality evaluation

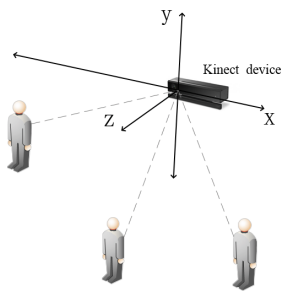

Fig. 5. The angle of the human body under the Kinect perspec-tive

There are two image quality evaluation in the general idea. One is subjective evaluation: subjective evaluation of the image quality by the observer; the other is objective evaluation: the use of algorithms to assess the image quality. Subjective evaluation methods are unavoidable and consistent with the subjective feelings, but they are also subjects to the subjective factors such as the professional background, psychological and mood of the observer. Objective evaluation

method is accurate and fast. It has a unique assessment value and is more suitable for use in the actual project, but it and the subjective feelings of people have a certain access.

This paper uses a gradient-based common image sharpness evaluation function Tenengrad [8]. The function employs operators to extract gradient values in both horizontal and vertical directions respectively, Defined as followed

$$Ten = \frac{1}{n} * \sum_{x} \sum_{y} S(x,y)^2 \tag{3}$$

Where $S(x,y) = \sqrt{G_x * I(x,y) + G_y * I(x,y)}$ is the gradient of the image I at $(x,y), G_x, G_y$ for *Sobel* convolutions, Respectively as

$$G_x = \begin{bmatrix} -1\ 0\ +1 \\ -2\ 0\ +2 \\ -1\ 0\ +1 \end{bmatrix}, G_y = \begin{bmatrix} -1\ -2\ -1 \\ 0\ \ 0\ \ 0 \\ +1\ +2\ +1 \end{bmatrix} \tag{4}$$

n is the total number of pixels in the image.

Based on the above indicators and assigned different weights, the overall score for each position is K

$$K = \frac{1}{2}\Delta + \frac{3}{10}Ten + \frac{1}{5}\chi \tag{5}$$

Each site of the human body takes the highest rated image as a sample.

3 Feature Design

The process of human re-recognition in this article is shown in Fig. 6.

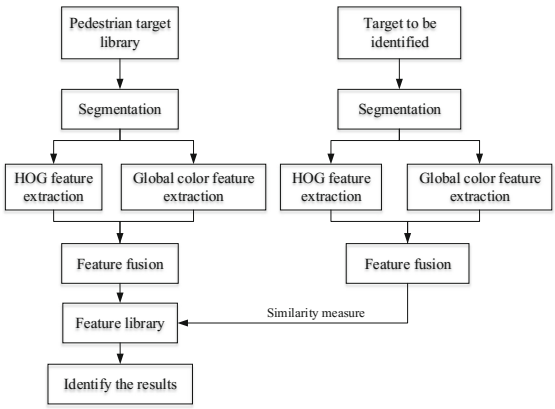

Fig. 6. Receptor flow chart of human target

3.1 Global Color Characteristics

The color feature is an important attribute to the image. RGB color space is composed of three primary colors that are mixed, the physical meaning is clear and suitable for the work of the picture tube, but the RGB color space does not have the light either the deformation and spectral invariance, and HSV color space is relative to these advantages. So the database image from RGB space to HSV space to experiment. Although the color histogram have the advantages which are simple to calculate, it has features that are insensitive to scale. But loses the spatial relationship between colors. Aibing Rao proposed the use of circular color histogram [9] to characterize the spatial characteristics of the color of the method to solve the problem.

Make $A_{ij} = |R_{ij}|$, for $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$, so get an matrix $M \times N : A = (A_{ij})_{M \times N}$. This matrix is a circular color histogram that represents the number of colors in a ring, the row represents the color value, the column represents the number of rings, $|R_{ij}|$ indicates the number of color values i in the j rings. The measure of the circular color histogram is defined as

$$d(I, J) = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (A_{ij} - B_{ij})^2} \tag{6}$$

3.2 Directional Gradient Histogram

The so-called histogram of oriented gradients (HOG) refers to a local region descriptor based on the gradient direction [10]. It constructs local and global surface features by statistically localized gradient histograms. Due to the normalization of local and global gradient histograms, it has a strong anti-interference ability for slight changes in the surface of the human body caused by changes in light factors. The main idea of the HOG feature is to describe the shape of the local target in the foreground of the human body using the gradient of the gradient or edge.

Set the gradient of pixel (x, y) in the input image be

$$G_x(x, y) = H(x + 1, y) - H(x - 1, y) \tag{7}$$

$$G_y(x, y) = H(x, y + 1) - H(x, y - 1) \tag{8}$$

Where $G_x(x, y)$, $G_y(x, y)$, $H(x, y)$ respectively represent the horizontal gradient, vertical gradient, and pixel values at pixel (x, y) in the input image. The gradient amplitude $G(x, y)$ and the gradient direction $\alpha(x, y)$ at pixel (x, y) are

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{9}$$

$$\alpha(x, y) = \tan^{-1}[\frac{G_x(x, y)}{G_y(x, y)}] \tag{10}$$

3.3 Similarity Measure

Bhattacharyya distance has been widely used in image processing and computer vision due to the faster speed of operation. But because of the real scene at different times, the location of the same human body targets often exist posture, angle changes, easily lead to HOG characteristics of the shift. If the use of only two goals corresponding to the gradient histogram comparison of the Bhattacharyya distance method, the mismatch rate is bound to increase. The EMD cross distance [11] avoids this situation, which is often used to measure the similarity of a set.

The HOG feature P of the human body is represented as a set of multiple feature sets, $P = ((\alpha_1, \omega_{\alpha 1}), (\alpha_2, \omega_{\alpha 2}), \ldots, (\alpha_m, \omega_{\alpha m}))$, α_i represents a directional gradient histogram vector, and $\omega_{\alpha i}$ represents the weight of the vector α_i. Then the EMD distance of the HOG feature $P_A = ((a_1, \omega_{a1}), (a_2, \omega_{a2}), \ldots, (a_m, \omega_{am}))$ of the target A and the HOG feature $P_B = ((b_1, \omega_{b1}), (b_2, \omega_{b2}), \ldots, (b_m, \omega_{bm}))$ of the target B are defined as

$$D_{EMD}(A, B) = \min_{f_{ij}} \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}, i = 1 \ldots m; j = 1 \ldots n \tag{11}$$

Where d_{ij} is the European distance of vector a_i and vector b_j, and f_{ij} is the transport stream.

Fusing the color feature and the HOG feature matching result, and then measuring the similarity among all the human body in the target body and the candidate set. Set γ, μ be the weight of the color feature and HOG, and the integrated distance between the two targets A and B is

$$D(A, B) = \gamma D_{color}(A, B) + \mu D_{EMD}(A, B) \tag{12}$$

Where $D_{color}(A, B) = \sqrt{1 - \sum_i \frac{I_A(i) \cdot I_B(i)}{\sum_i I_A(i) \cdot I_B(i)}}$.

Several experiments were done to set the weight to $\gamma = 0.65, \mu = 0.35$.

4 Experimental Results and Analysis

Generally speaking, the body re-recognition will have two sets of data sets, query set and candidate set. This experiment is carried out in the Kinect REID database and the BIWI RGBD-ID database. These two databases are based on the depth of information on the human body to re-identify the common database. General re-recognition of the human body as a similarity sorting problem. The mainstream of target identification criteria is CMC curve(cumulative matching curve) [12]. The ranking k in the CMC curve represents the search for the target body in the candidate set. In the first k search. Results find the ratio of the target body to be queried. In this paper, the use of the evaluation criteria, and only use of color features, only use of HOG features and projection method of three methods for comparison.

4.1 Test Results and Analysis of Database Kinect REID

Kinect REID database have total of 71 human bodies, each body have about 120 frames, Fig. 7 are parts of the data example. 120 frames data of each human body randomly selected 60 as the query set, the remaining 60 frames as a candidate set, the two sets of each human body target segmentation, and then select the optimal score through the part of the frame.

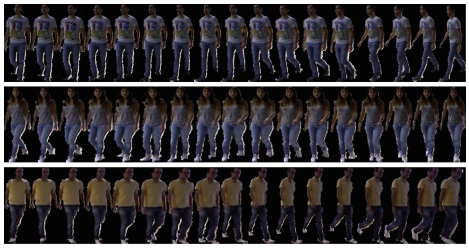

Fig. 7. Parts of the data in the database Kinect REID

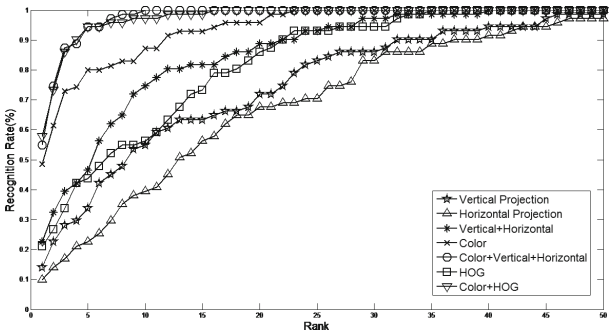

Fig. 8. The algorithm results in the database Kinect REID

As can be seen from Fig. 8 combining the color feature with the HOG feature significantly improves the recognition rate. Table 1 shows the comparison of the results of this algorithm with the SDALF, MCMimpl, SGLTrP3 and ED + SKL methods. It can be seen that the SGLTrP3 recognition rate is 66%, better than this and other methods at the first matching rate. But from the fifth recognition rate, this method is 94.3%, much higher than the other four methods.

4.2 Test Results and Analysis of Database BIWI RGBD-ID

Database BIWI RGBD-ID consists of two parts. Namely, Still dataset and Walking dataset, each part also contains training set of 28 human bodies and test set

Table 1. The algorithm is compared with other algorithms in the Kinect REID database

Method	Rec[a] rate 1	Rec rate 5	Rec rate 10	Rec rate 30	Rec rate 50
SDALF [13]	41%	70%	82%	98%	100%
MCMimpl [13]	51%	78%	87%	99%	100%
SGLTrP3 [14]	66%	82%	91%	100%	100%
ED + SKL [15]	56%	86%	94%	100%	100%
The method of this article	57.7%	94.3%	97.2%	100%	100%

[a]Notes: Rec is logogram of Recognition.

Table 2. The algorithm in this section is compared with other algorithms in the Still dataset

Method	Rec[a] rate 1	Rec rate 5	Rec rate 10	Rec rate 30	Rec rate 50
Face + Skeleton (SVM) [16]	52%	81%	90%	98%	100%
Nearest Neighbor [17]	27%	45%	82%	97%	100%
ED + SKL [15]	31%	68%	81%	100%	100%
The method of this article	58.9%	91.0%	98.7%	100%	100%

[a]Notes: Rec is logogram of Recognition.

of 50 human bodies. Taking into account the lack of training part of this article and if only a separate use of the query set or candidate set of experimental samples less, so mix the training set and test set together, there are total 78 people for re-recognition experiments (Table 2).

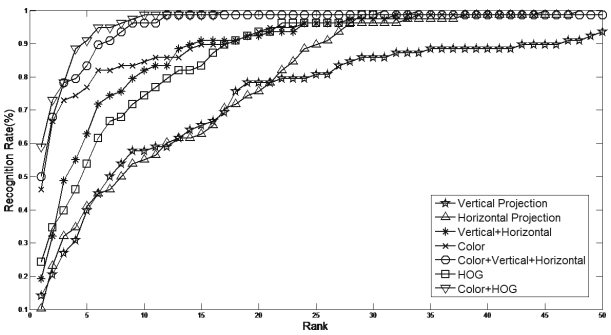

Fig. 9. The experimental results of this algorithm in the Still dataset section

Figure 9 shows the recognition results of the seven curves using the color feature, the typical method, the HOG feature, and the combination of the two in the Still dataset section. When the HOG feature is used only, the recognition rate of the top 10 is lower, but after the combination of the two of then, the

recognition rate has been significantly improved, the first matching rate reach at 58.9%, the 10th matching rate at 98.7% which is higher than other similar algorithms (Fig. 10).

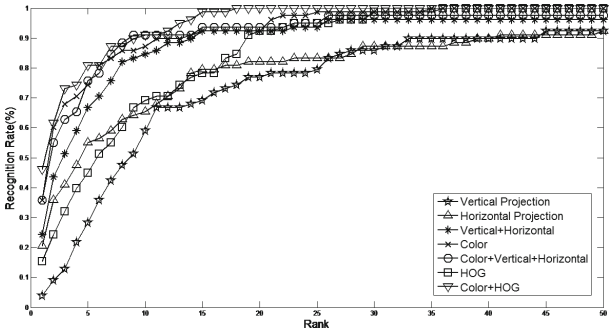

Fig. 10. The experimental results of this algorithm in the Walking dataset section

Table 3. The algorithm in this section is compared with other algorithms in the Walking dataset

Method	Rec[a] rate 1	Rec rate 5	Rec rate 10	Rec rate 30	Rec rate 50
Face + Skeleton (SVM) [16]	43.9%	74%	85%	96.5%	100%
Nearest Neighbor [17]	21%	43%	77%	97%	100%
ED + SKL [15]	26%	61%	78%	100%	100%
The method of this article	46.2%	80.7%	91.0%	100%	100%

[a]Notes: Rec is logogram of Recognition.

Table 3 compares this algorithm with the methods of machine learning algorithms such as Face + Skeleton (SVM), Nearest Neighbor and ED + SKL, except that the method is slightly 2.3% higher than the SVM method at the first matching rate. The rate of the beginning of the other matching rate, the algorithm has achieved the highest recognition rate.

5 Conclusion

In this paper, we propose an accurate segmentation of the human body according to the depth information of the human joints. The optimal frame is selected as the experimental sample for the multi-frame images of a human. It is proved that the human body segmentation based on deep bone nodes can enhance the robustness to occlusion problem, angle of view change and attitude change. In the re-recognition stage, a new object recognition algorithm combining color feature and HOG feature is proposed to improve the recognition rate. The next step is to improve the HOG characteristics and find better description features to further improve the recognition rate.

References

1. Farenzena, M., Bazzani, L., Perina, A., et al.: Person re-identification by symmetry-driven accumulation of local features. In: Computer Vision and Pattern Recognition, pp. 2360–2367. IEEE (2010)
2. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: what features are important? In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 391–401. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_39
3. Qin, H.K., et al.: Summary of intelligent video surveillance technology. J. Comput. Sci. **38**(6), 1093–1118 (2015). In chinese
4. Alahi, A., Vandergheynst, P., Bierlaire, M., et al.: Cascade of descriptors to detect and track objects across any network of cameras. Comput. Vis. Image Underst. **114**(6), 624–640 (2010)
5. Alahi, A., Marimon, D., Bierlaire, M., et al.: A master-slave approach for object detection and matching with fixed and mobile cameras. In: IEEE International Conference on Image Processing, pp. 1712–1715. IEEE (2008)
6. Bak, S., Corvee, E., Brmond, F., et al.: Person re-identification using spatial covariance regions of human body parts. In: IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 435–440. IEEE (2010)
7. Shotton, J., Kipman, A., Kipman, A., et al.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**(1), 116–124 (2013)
8. Arun, R., Nair, M.S., Vrinthavani, R., et al.: An alpha rooting based hybrid technique for image enhancement. Eng. Lett. **19**(3), 159–168 (2011)
9. Rao, A., Srihari, R.K., Zhang, Z.: Spatial color histograms for content-based image retrieval. In: IEEE International Conference on TOOLS with Artificial Intelligence, p. 183. IEEE Computer Society (1999)
10. Dalal, N., Triggs, B., Triggs, B.: Histograms of oriented gradients for human detection. CVPR **1**(12), 886–893 (2005)
11. Fu, A.Y., Liu, W., Deng, X.: Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD). IEEE Trans. Dependable Secure Comput. **3**(4), 301–311 (2006)
12. Bolle, R.M., Connell, J.H., Pankanti, S., et al.: The Relation between the ROC curve and the CMC. In: Fourth IEEE Workshop on Automatic Identification Advanced Technologies, vol. 2005, pp. 15–20. IEEE (2005)
13. Pala, F., Satta, R., Fumera, G., et al.: Multimodal person reidentification using RGB-D cameras. IEEE Trans. Circ. Syst. Video Technol. **26**(4), 788–799 (2016)
14. Imani, Z., Soltanizadeh, H.: Person reidentification using local pattern descriptors and anthropometric measures from videos of kinect sensor. IEEE Sens. J. **16**(16), 6227–6238 (2016)
15. Wu, A., Zheng, W.S., Lai, J.H.: Robust Depth-Based Person Re-Identification. IEEE Press (2017)
16. Munaro, M., Basso, A., Fossati, A., et al.: 3D reconstruction of freely moving persons for re-identification with a depth sensor. In: IEEE International Conference on Robotics and Automation, pp. 4512–4519. IEEE (2014)
17. Munaro, M., Fossati, A., Basso, A., Menegatti, E., Van Gool, L.: One-shot person re-identification with a consumer depth camera. In: Gong, S., Cristani, M., Yan, S., Loy, C.C. (eds.) Person Re-Identification. ACVPR, pp. 161–181. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6296-4_8