IPSJ Transactions on Computer Vision and Applications Vol.5 94–98 (July 2013)

[DOI: 10.2197/ipsjtcva.5.94]

Express Paper

Depth-based Gait Feature Representation

Hozuma Nakajima[1,a] Ikuhisa Mitsugami[1,b] Yasushi Yagi[1,c]

Received: March 11, 2013, Accepted: April 24, 2013, Released: July 29, 2013

Abstract: This paper proposes a novel gait feature representation that well describes characteristics of a walking person from the perspective of a range sensor. Most existing methods for gait feature extraction use a sequence of his/her silhouette as their input, so that they inevitably suffer from the difficulty of silhouette extraction in real scenes and change of view direction, which prevent them from being applied in practice. The proposed method, on the other hand, does not require such accurate segmentation, and is not affected by view change since captured range data has three-dimensional information. In addition, our method can explicitly separate dynamic feature from a static one, e.g., body shape, which have never been realized. Experimental results of gait authentication show its effectiveness.

Keywords: gait, feature extraction, range data, authentication

1. Introduction

Walking is a fundamental behavior of our daily lives. Gait (i.e., way of walking) thus gets more attention as a useful biometric in many fields: person authentication [1], [2], [3], [4], [5], [6], [7], age estimation [8], gender estimation [9], etc. Performance reported in these studies are in fact good enough to be applied in the real-world. They have, however, never been applied yet. The essential reason is that their features are based on silhouettes of a walking person so that they need accurate silhouettes, which is usually difficult to obtain in real scenes. There are also approaches that use other features instead of the silhouettes [10], [11], their performance is still worse than those by the silhouette-based methods. Change of view direction is also another problem. The silhouette-based methods implicitly assume the person is captured by an orthogonal camera, so that as long as we use a practical camera, which is modeled as a projective camera, his/her silhouette is distorted according to his/her position in captured images. These problems, the difficulty of silhouette extraction and view change, prevent the gait analysis techniques from being applied in practice.

Motivated by this fact, this paper proposes a novel gait feature representation method that well describes characteristics of a walking person from the perspective of a range sensor. We call this feature Depth-based Gait Feature (DGF). The proposed method does not require accurate segmentation; we need just the position of the person and then use depth data around the position. In addition, our method is not affected by view change since the depth data can be aligned correctly by considering the view change. This should be the first method for gait representation that can be applied to real scenes.

Our method has another interesting property that it can explicitly separate motion and shape features, which has never been achieved in other studies. The direct component and the amplitude/phase components of DGF correspond to the shape and motion features, respectively. Sivapalan et al. [12] also propose a depth-based feature, but it is just a simple extension of Gait Energy Image (GEI) [2] and thus shape and motion features are not separated.

To evaluate the ability of DGF, we experimentally apply it to person authentication task. The experimental results show that our method gives better performance than a state-of-the-art 2-D silhouette-based method.

2. Algorithm

2.1 Pedestrian Detection and Tracking

Figure 1 (a) shows an example of depth image as captured by a range sensor. We assume the range sensor is preliminarily calibrated so that the depth image can be transformed into a three-dimensional (3-D) point cloud in the world coordinate system, and vice versa. A walking person is detected by background subtraction. Note that we do not need to extract his/her accurate silhouette. His/her position in the world coordinate system (X_s, Y_s)

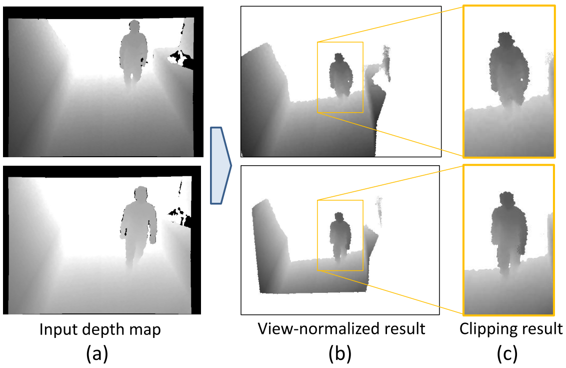


Fig. 1 View normalization and clipping to obtain GDS.

[1] The Institute of Scientific and Industrial Research, Osaka University, Ibaraki, Osaka 567–0047, Japan
[a] nakajima@am.sanken.osaka-u.ac.jp
[b] mitsugami@am.sanken.osaka-u.ac.jp
[c] yagi@am.sanken.osaka-u.ac.jp

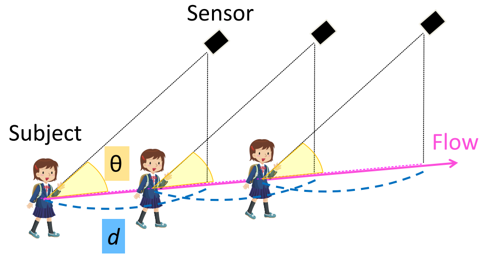


Fig. 2 Virtual viewpoints.

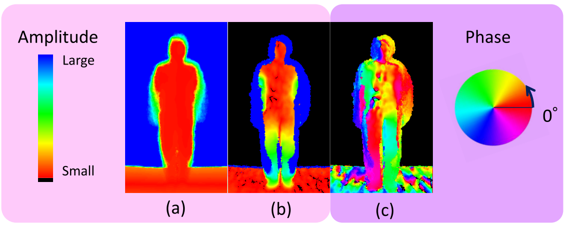


Fig. 3 Amplitude spectrum and phase.
(a) $A(x, y, 0)$, (b) $A(x, y, 1)$, (c) $\Theta(x, y, 1)$.

is calculated as the gravity point of three-dimensional points corresponding with the region. The person is then tracked simply by applying this detection process to each frame.

2.2 View Normalization and Clipping

In the case of the two-dimensional (2-D) silhouette-based methods, silhouettes in a sequence are aligned so that their gravity points should be constant, to calculate an average image in the case of GEI [2] or apply Fourier transformation in Ref. [3]. Since there is no 3-D information, however, they suffer from the fact that the silhouette is distorted according to the position in a captured image caused by perspective effect of cameras, which is inevitable even when the person walks straight and the camera is located perpendicularly to his/her walking direction. On the other hand, in the case of our study, we have 3-D information of the person so that we can overcome this distortion problem by generating the depth image of a virtual viewpoints that is located at a relatively constant position from the person.

Figure 2 describes how to determine the virtual viewpoint. We first obtain the person's direction $\boldsymbol{v} = (v_x, v_y)$ ($|\boldsymbol{v}| = 1$) by differentiating his/her trajectory. The position (X_c, Y_c, Z_c) and orientation $(\phi_{Xc}, \phi_{Yc}, \phi_{Zc})$ is then calculated as follows:

$$(X_c, Y_c, Z_c) = (X_s - v_x d, Y_s - v_y d, d\tan(\theta) + Z), \tag{1}$$

$$(\phi_{Xc}, \phi_{Yc}, \phi_{Zc}) = (0, \theta, \tan^{-1}(v_y/v_x)), \tag{2}$$

where θ denotes the depression angle of the range sensor. Figure 1 (b) shows generated depth images using the extrinsic parameters. Since the virtual camera follows the person, he/she appears always at the center of the depth image regardless of where he/she is. We are thus able to obtain a sequence of the depth image around the person just by clipping a window and resizing it to a fixed size $H \times W$, as shown in Fig. 1 (c). We call the sequence of the clipped depth images as a Gait Depth-map Sequence (GDS). Note that we do not care if each pixel is part of the foreground or background at this step. GDS thus also contains floor regions around his/her feet.

2.3 Frequency Domain Feature

By considering the periodicity of walking, we first estimate a period N_{gait} and extract frames corresponding with a cycle from GDS by evaluating its autocorrelation, which is a similar way to Fourier Domain Feature (FDF) [3]. We then apply Discrete Fourier Transformation (DFT) to the extracted frames.

$$G(x, y, k) = \sum_{n=0}^{N_{gait}-1} g(x, y, n) e^{-j\omega_0 kn}, \tag{3}$$

where $g(x, y, n)$ denotes a depth value at a pixel (x, y) in the n-th frame, ω_0 is a base angular frequency for the gait period N_{gait}, and $G(x, y, k)$ is the DFT of GDS for k-times the gait period. Then, an amplitude spectrum $A(x, y, k)$ is calculated as follows:

$$A(x, y, k) = \frac{1}{N_{gait}} |G(x, y, k)|. \tag{4}$$

In most existing studies, they use only the amplitude information for gait analysis. In this study, however, we also focus on phase information. This phase component $\Theta(x, y, k)$ is calculated as follows:

$$\Theta(x, y, k) = \begin{cases} \tan^{-1}\left(\dfrac{\mathrm{Im}[G(x, y, k)]}{\mathrm{Re}[G(x, y, k)]}\right) & (\mathrm{Re}[G(x, y, k)] > 0), \\ \pi + \tan^{-1}\left(\dfrac{\mathrm{Im}[G(x, y, k)]}{\mathrm{Re}[G(x, y, k)]}\right) & \text{otherwise}, \end{cases} \tag{5}$$

where $\mathrm{Re}[G(x, y, k)]$ and $\mathrm{Im}[G(x, y, k)]$ are the real and imaginary parts of $G(x, y, k)$, respectively. **Figure 3** shows examples of $A(x, y, 0)$, $A(x, y, 1)$, and $\Theta(x, y, 1)$. The direct component $A(x, y, 0)$ mainly describes the shape of the person. On the other hand, $A(x, y, 1)$ and $\Theta(x, y, 1)$ correspond to the fundamental motion of walking such as arm swings and steps that occur only once in a period, so that pixel values around the arms and legs are higher than ones around the trunk in $A(x, y, 1)$ and the right arm and left leg have similar phase while the left and right arms have the opposite phase.

$A(x, y, k), \Theta(x, y, k)$ ($k \geq 2$) correspond with amplitudes and phases of k-times the frequency. These are, however, expected to be noisy and less reliable considering the number of frames included in a cycle and the noise of each range sensor. In this paper, therefore, they are not used for analysis.

2.4 Thresholding

In Fig. 3 (b), there are blue regions, which mean much higher values, around the person's edge in $A(x, y, 1)$. This is because of a pixel in the region either on the human body or on the background because of his/her fluctuation. Shapes of these regions have information about his/her walking, but they are actually described also in $A(x, y, 0)$ (as green regions). To reduce the redundancy, the blue regions are eliminated using a threshold τ as follows:

$$A'(x, y, 1) = \begin{cases} A(x, y, 1) & (A(x, y, 1) > \tau), \\ Nan & \text{otherwise}. \end{cases} \tag{6}$$

Note that this thresholding is not sensitive at all. The depth of the background is much larger than that of the range of the human shape and motion.

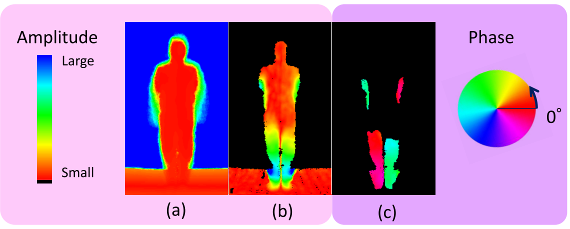


Fig. 4 Amplitude spectrum and phase after thresholding. (a) $A(x, y, 0)$, (b) $A'(x, y, 1)$, (c) $\Theta'(x, y, 1)$.

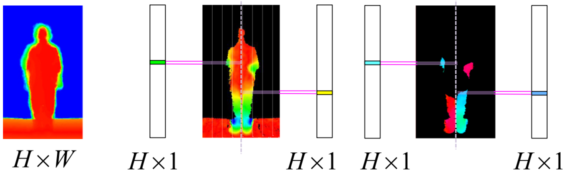


Fig. 5 Feature representation of our method.

As for the phase component $\Theta(x, y, 1)$, we find that pixel values around the trunk is very noisy and unstable. This is because the range of depth change around the trunk is so small compared with the measurement accuracy of range sensors. Considering this fact, we also eliminate the corresponding regions by masking $\Theta(x, y, 1)$ by $A'(x, y, 1)$. In addition, considering that the extracted periods do not begin with a constant phase walking, a certain pixel (x_0, y_0) (on the right leg) is determined as a reference pixel, and all pixel values are normalized by the value of the reference pixel. This procedure is described as follows:

$$\Theta'(x, y, 1) = \begin{cases} \Theta(x, y, 1) - \Theta(x_0, y_0, 1) & (A'(x, y, 1) \neq Nan) \\ Nan & (A'(x, y, 1) = Nan) \end{cases} \tag{7}$$

Figure 4 shows $A'(x, y, 1)$ and $\Theta'(x, y, 1)$. These images describe well the fundamental motion of walking.

2.5 Feature Representation

Comparing $A'(x, y, 1)$ and $\Theta'(x, y, 1)$ with $A(x, y, 0)$ in Fig. 4, we find they still contain redundancy; $A'(x, y, 1)$ and $\Theta'(x, y, 1)$ are strongly affected by the human shape that is described in $A(x, y, 0)$. To reduce the redundancy, as shown in **Fig. 5** we compress them into one-dimensional vectors a and t by choosing the maximum value for each row as follows:

$$\boldsymbol{a}_R = \{a_{Rj} = \max_{0 \leq i \leq \frac{W}{2}} A'(i, j, 1)\} \ \ (i = 1, \cdots, H) \tag{8}$$

$$\boldsymbol{a}_L = \{a_{Lj} = \max_{\frac{W}{2} \leq i \leq W} A'(i, j, 1)\} \ \ (i = 1, \cdots, H) \tag{9}$$

$$\boldsymbol{t}_R = \{t_{Rj} = \max_{0 \leq i \leq \frac{W}{2}} \Theta'(i, j, 1)\} \ \ (i = 1, \cdots, H) \tag{10}$$

$$\boldsymbol{t}_L = \{t_{Lj} = \max_{\frac{W}{2} \leq i \leq W} \Theta'(i, j, 1)\} \ \ (i = 1, \cdots, H) \tag{11}$$

Finally, the gait feature in this paper is defined as a vector consisting of the direct component $A(x, y, 0)$, the amplitude component $\boldsymbol{a} = \{\boldsymbol{a}_R, \boldsymbol{a}_L\}$, and the phase component $\boldsymbol{t} = \{\boldsymbol{t}_R, \boldsymbol{t}_L\}$. We call this new gait feature as Depth-based Gait Feature (DGF). The number of dimension of this feature is thus $H \times (W + 4)$.

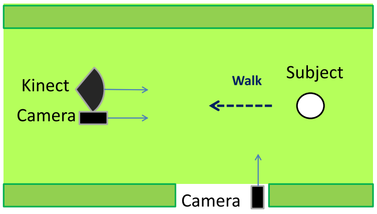
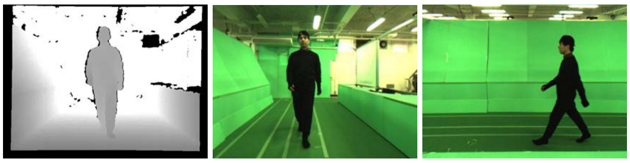


Fig. 6 Experimental setting.



(a) Kinect (front) (b) Camera (front) (c) Camera (side)

Fig. 7 Captured images in experimental environment.

3. Experiment

3.1 Settings

To evaluate the effectiveness of this feature representation, we apply the feature to a person authentication task, which is in fact the main interest in the gait analysis field.

Figure 6 shows our experimental environment. There is a straight walking path. Microsoft Kinect, which is used as a range sensor, is located in front of the path. In addition, we also locate two conventional cameras; one is located at the same position as the Kinect, and the other is located to observe a walking person from his/her side. **Figure 7** shows examples of their captured images. There are 31 subjects, and there are five sequences for each subject. A sequence is for the gallery and the others are used as probes. Note that since all walls and floor are colored green and we asked each subject to wear a black suit that covers the entire body, their silhouette can be easily and accurately obtained from images captured by the cameras. As a result, the quality of the silhouettes is high enough. On the other hand, as the Kinect cannot measure depths around occluding edges, the edges of the subjects in the depth images are much less clear than those of the cameras.

For the similarity measurement of features, we use the *L2* norm of each component. For this calculation, pixels labeled as *Nan* are ignored.

3.2 Comparison with a Silhouette-based Method

For comparative evaluation, we employed ROC curves to indicate the trade-offs between the false rejection rate (FRR) of the genuine and the false acceptance rate (FAR) of the imposter while changing the acceptance threshold. **Figure 8** shows the ROC curves: (a) and (b) are the curves using FDF [3] as obtained from camera images, and (c) is from using DGF as obtained from depth images. From this result, it is confirmed that DGF gives better performance than FDF.

In addition, to clarify the reason of the performance, Fig. 8 also shows the curve of DGF with height normalization, which means that the height cue is eliminated from the feature. From this result, we confirmed that even when there is no height cue, our depth-based feature still gives better performance than FDF. This

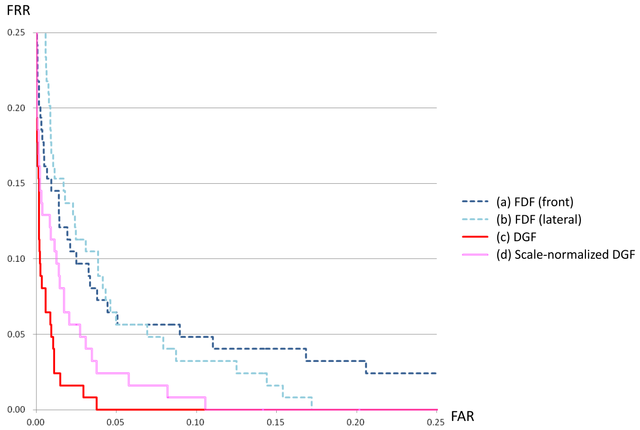


Fig. 8 Effect of scale normalization.

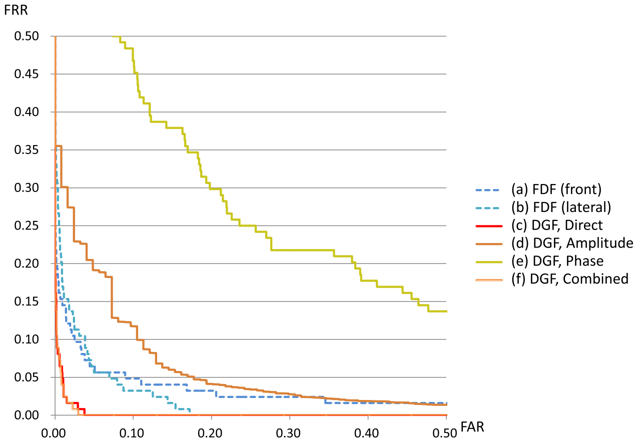


Fig. 9 Comparison between a 2-D silhouette based method and our proposed method (ROC curves).

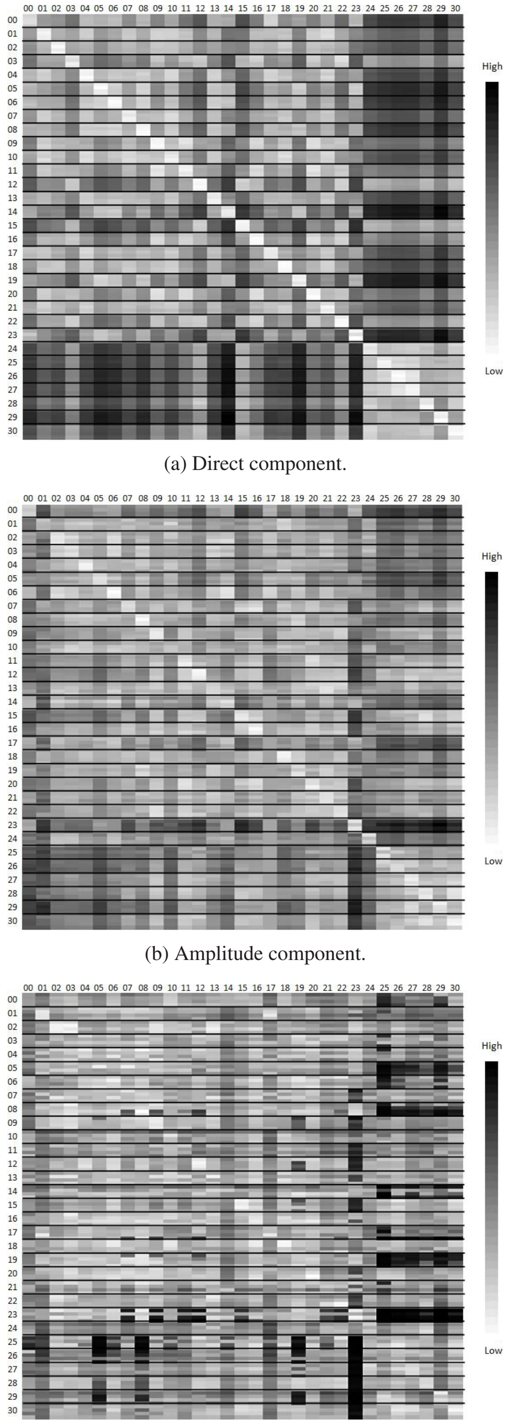
Table 1 Performance comparison by equal error rate (EER).

Feature	EER [%]
FDF (fron)	5.64
FDF (lateral)	5.64
DGF, Direct	1.61
DGF, Amplitude	10.48
DGF, Phase	25.00
DGF, Combined	1.61

fact indicates that the 2-D silhouette-based method is inevitably affected by the perspective distortion, while our 3-D depth-based proposed method is not affected since it can utilize 3-D information.

3.3 Comparison of Components

Figure 9 shows how each component in our method contributes to the performance of the person authentication task. An equal error rate (EER) of the FAR and FRR is another measurement for comparison. **Table 1** shows EERs of these components. In addition, **Fig. 10** shows their confusion matrices. In each matrix, a row and column correspond with gallery and probe, respectively, and there are five rows for each subject.



(a) Direct component.

(b) Amplitude component.

(c) Phase component.

Fig. 10 Confusion matrix of each component.

In Fig. 9, the direct component (c) gives much better performance than the amplitude component (d) and the phase component (e). The performance when combining all of the components (f) is almost the same as that of the direct component only (c); the direct component is dominant for this task. The amplitude component, however, also has quite a high discrimination ability, considering its number of dimension is much lower than the direct one and it is only motion information without any shape information. As for the phase component, it seems not as effective as long as it is used for the person authentication task. It is in fact intuitively understandable that the phase is not so person-oriented; everyone swings their right arms and left legs in similar phase and their left and right arms in opposite phase. It is, nevertheless, notable that this is the first feature representation method that extracts such phase components so clearly. We expect the feature would be effective for, e.g., the detailed analysis of slight gait change caused by different clothes, shoes, emotion, etc.

4. Conclusion

This paper proposes a novel gait feature representation method that well describes the characteristics of a walking person from the perspective of a range sensor. This proposed method does not require accurate segmentation, and is not affected by view change since the captured range data has three-dimensional information. These characteristics allow the method to be applied to general scenes; we do not need to worry about the difficulty of silhouette extraction, nor locate a camera at a distant position from a person to be able to assume orthogonal projection, which is required for existing 2-D silhouette-based methods. It is also a notable advantage of our method that it can explicitly separate motion and shape features, which has never been realized in other studies. We apply the proposed method to the person authentication task to evaluate its effectiveness. The experimental results says that our method gives better performance than a state-of-the-art 2-D silhouette-based method, though the direct component is dominant for this task.

Limitations. Our method assumes that a range sensor captures a person roughly from his/her front. As long as the change of his/her direction is not so large, the method works well thanks to view normalization. If the change is large, however, it fails because GDS contains many *Nan* pixels.

Acknowledgments This work was supported by CREST (Core Research for Evaluation Science and Technology) of JST (Japan Science and Technology).

References

[1] Lynnerup, N. and Vedel, J.: Person identification by gait analysis and photogrammetry, *Journal of Forensic Science*, Vol.50, No.1, pp.112–118 (2005).

[2] Han, J. and Bhanu, B.: Individual recognition using gait energy image, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, pp.316–322 (2006).

[3] Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T. and Yagi, Y.: Gait recognition using a view transformation model in the frequency domain, *Proc. 9th European Conference on Computer Vision*, pp.151–163 (2006).

[4] Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T. and Yagi, Y.: Adaptation to walking direction changes for gait identification, *IEEE International Conference on Pattern Recognition*, Vol.2, pp.96–99 (2006).

[5] Larsen, P.K., Simonsen, E.B. and Lynnerup, N.: Gait analysis in forensic medicine, *Journal of Forensic Sciences*, Vol.53, No.5, pp.1149–1153 (2008).

[6] Lam, T.H.W., Cheung, K.H. and Liu, J.N.K.: Gait flow image: A silhouette-based gait representation for human identification, *Pattern Recognition*, Vol.44, pp.973–987 (2011).

[7] Bouchrika, I., Goffredo, M., Carter, J. and Nixon, M.: On using gait in forensic biometrics, *Journal of Forensic Science*, Vol.56, No.4, pp.882–889 (2011).

[8] Mannami, H., Makihara, Y. and Yagi, Y.: Gait analysis of gender and age using a large-scale multi-view gait database, *Proc. 10th Asian. Conf. Computer Vision*, pp.975–986 (Nov. 2010).

[9] Iwama, H., Makihara, Y., Okumura, M. and Yagi, Y.: Gait-based age estimation using a whole-generation gait database, *Proc. International Joint Conference on Biometrics* (*IJCB2011*), pp.1–6 (2011).

[10] Wang, X., Doretto, G., Sebastian, T., Rittshcer, J. and Tu, P.: Shape and appearance context modeling, *Proc. 11th International Conference on Computer Vision*, pp.1–8 (2007).

[11] Kawai, R., Makihara, Y., Hua, C., Iwama, H. and Yagi, Y.: Person Re-identification using View-dependent Score-level Fusion of Gait and Color Features, *Proc. 21st International Conference on Pattern Recognition* (*ICPR2012*), pp.2694–2697 (2012).

[12] Sivapalan, S., Chen, D., Denman, S., Sridharan, S. and Fookes, C.: Gait energy volumes and frontal gait recognition using depth images, *Proc. International Joint Conference on Biometrics* (*IJCB2011*), pp.1–6 (2011).

(Communicated by *Tatsuya Harada*)