Human Gait Recognition Using Depth Camera: A Covariance Based Approach

Naresh Kumar M S
Video Analytics Lab, SERC
Indian Institute of Science
Bangalore, India
nareshkms88@gmail.com

R. Venkatesh Babu*
Video Analytics Lab, SERC
Indian Institute of Science
Bangalore, India
venky@serc.iisc.in

ABSTRACT

Gait is an important biometric modality for recognizing humans. Unlike other biometrics, human gait can be captured at a distance which makes it an unobtrusive method for recognition. In this paper, an unrestricted gait recognition algorithm is proposed which uses 3D skeleton information and trajectory covariance of joint points. 3-D skeleton is generated from the depth images that are captured using *Kinect* sensor. The temporal tracking of skeleton points is used for gait analysis. The covariance measure between these skeleton point trajectories are computed and the covariance matrices form the gait model. The gait is recognized by computing the minimum dissimilarity measure between the gait models of the training data and the testing data. Recognition accuracy of over 90% has been achieved for a data set consisting of fixed and moving camera scenarios of 20 subjects.

Keywords

Gait recognition, human recognition, depth sensor, human skeleton, covariance, covariance dissimilarity.

1. INTRODUCTION

Automatic gait recognition is an important biometric tool for recognizing people at a distance based on the individual's style of walking. Gait recognition tries to mimic the capability of humans to recognize people based on their walking style even at a long distance. Gait is an important modality for crucial applications such as video surveillance, due to it's surreptitiousness. Unlike other non-invasive biometrics such as face and iris recognition, gait does not require high resolution images or special equipments. In most of the algorithms, gait of the person is recognized when the person presents the side view.

*Corresponding author

ICVGIP '12, December 16-19, 2012, Mumbai, India

Gait recognition algorithms are broadly divided into two approaches: i) Model based and ii) Model free [1]. Model based approaches use the information from body parts such as joints for constructing a model. Though this approach is typically view invariant, it requires very high quality gait sequences. Such a system requires a multi-camera set-up for collecting the video sequences. Johnson [2], attached light bulbs to the human subject and tracked the movement of the light bulbs to capture the subject's motion. In another system a stick model is created from body contour for recognizing gait [3]. Johnson et al., [4] proposed a multi view gait recognition algorithm utilizing static body parameters such as height and distance between various body points. Most of the model-free approaches use binary silhouette information for recognition. Extracting silhouette involves the use of background modelling techniques. Murase and Sakai [5] proposed a parametric eigenspace technique for recognition. The input silhouette image sequence is projected on eigenspace in order to form a trajectory in eigenspace. Distance between this trajectory and the reference trajectory is used for classification. Lam et al., [6] proposed a gait recognition algorithm that fuses motion and static spatio-temporal templates of silhouette image sequences. Chen et al., [7] proposed a robust dynamic gait representation scheme, frame difference energy image (FDEI), to suppress the influence of silhouette incompleteness.

Due to the subtle nature of the problem, most of these approaches rely on special set-up such as multi-camera systems, high quality video or superior silhouette information for gait recognition. Due to these requirements, often it is very difficult to perform gait recognition in an unrestricted ambience. But today with the advancement of camera and video technology, it is possible to capture the depth image which gives us information in the 3rd dimension (depth) with which we can easily extract the silhouette of the subject very accurately. The recently introduced, inexpensive *Kinect* sensor provides the depth information along with RGB colour info. The work documented in [8], [9] and [10] make use of *Kinect* sensor.

Palanquin et al. [8] have proposed the use of *Kinect* sensor to perform gait recognition using Gait Energy Volume (GEV) generated from multi-view silhouettes and frontally acquired depth images. Acquiring multi-view images is an extra cost and the results are dependent on the pose of the subject while acquiring the depth images.

Ball et al. [9] directly use the skeleton point trajectories. Although the method proposed in [9] is unsupervised, walk cycles have to be manually segmented and extracted. The

data from only the lower part of the body is used for recognition. This method reports net recognition accuracy of 43.6% for just 4 subjects. In our work, recognition accuracy of over 90% is achieved with a larger dataset of 20 subjects. With the proposed method it is no longer necessary to segment and synchronize the walk sequences. It was also found that the upper-body, mainly the arms make a significant contribution to the recognition performance, which is not used in [9].

Pries et al. [10] claim recognition rate of over 90% with 9 subjects again using skeleton point trajectory. However, the disadvantage in this method is that the subject always has to walk parallel to the camera plane so that the height and length features are constant for each subject. Suppose the subject walks towards the camera or walks parallel but at a different distance from the camera, then the subject's height varies and this method would fail. It is also important to note that *Kinect SDK* is designed to recognize the frontal portion of the human body and model it as a skeleton. Suppose a subject is standing with his back towards the camera, he is still detected as facing towards the camera. In the scenario used in [10] the right side of the body is occluded and it is not modelled as accurately as the left side and so only the trajectories of the left portion of the skeleton are useful. The algorithm proposed in our work accommodates for changes in skeleton size as the subject moves towards the camera. The data used contains walk sequences of subjects walking straight and towards the camera such that a major frontal portion of the subject is visible to the camera. Although this might seem like a drawback that the entire frontal portion of the subject has to be visible, it is not. All the data is at our disposal for training and testing can be done with a smaller portion of the skeleton. The experimental results show that a subset of the skeletal points are sufficient to achieve a recognition rate of over 90%.

Our method which also utilizes the skeleton point trajectories is independent of the pose of the human in front of the camera, which gives an advantage over [8], [9] and [10]. Our intuition being that the relative movement of a subject's body parts with each other forms a principal component of his gait, we use covariance measures between the skeletal point trajectories to model the gait and to capture the relative movement of the skeletal points of each subject. Later, identification is done based on the dissimilarity measures between the test and trained gait models. The database of 20 subjects were collected in both fixed and moving camera scenarios. The proposed approach attains more than 90% recognition rates even when one scenario is used for training and the other for testing. This shows the major advantage of the proposed approach for identifying humans in unrestricted indoor ambience.

The paper is organized as follows. Section 2 presents the overview of the proposed gait recognition system. Section 3 details the steps involved in the proposed approach for gait recognition. Section 4 discusses the experimental tests done and their results. The final concluding remarks are given in section 5.

2. SYSTEM OVERVIEW

The overview of the proposed system is illustrated in Figure 1. The Kinect sensor consists of an RGB camera and an infra-red structured light source-sensor combination for inferring depths. Depth images of resolution 640x480 are obtained in the form of grey-scale images, where the grey value is proportional to the distance of the object from the sensor scaled in the range of 1.2m to 3.5m. Skeleton from the depth images is generated using Microsoft's SDK for Kinect [11]. Generated skeleton consists of 20 points that are essential for modelling the human locomotion which are listed in Table 1. Figure 2 shows the 20 skeletal points along with their labels. The 3D trajectory information of these 20 points are recorded. Figure 3 shows the plots of the walk sequences for first scenario and second scenario. For the second scenario the irregularity can be clearly observed because of the camera movement. These trajectories are further processed to normalize the range of values of the trajectories and the skeleton size. The location of the skeleton is then offset with respect to the hip centre. Now the covariance between the points and the dissimilarity measure between the covariance matrices are computed to identify the subject as described in Sec. 3.

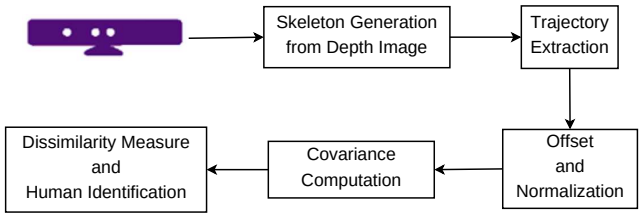

Figure 1: System Overview

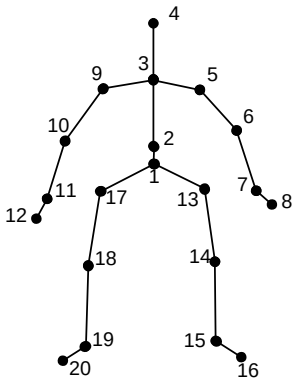

Figure 2: Skeleton with 20 joint points.

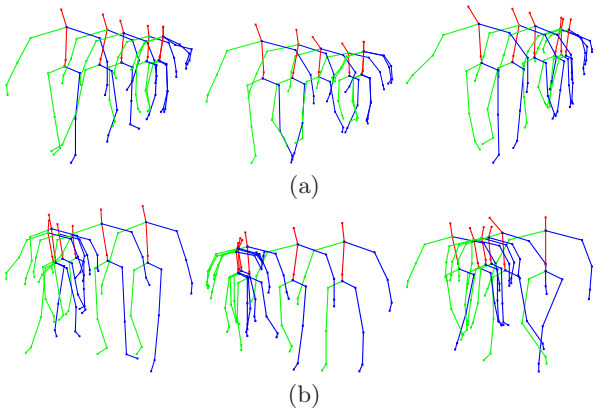

Figure 3: Skeletal plot for (a) Fixed camera (First scenario) (b)Freely moving camera (Second scenario).

Table 1: Skeleton points

Joint Number	Joint Name	Joint Number	Joint Name
1	Hip Centre	11	Wrist Right
2	Spine	12	Hand Right
3	Shoulder Centre	13	Hip Left
4	Head	14	Knee Left
5	Shoulder Left	15	Ankle Left
6	Elbow Left	16	Foot Left
7	Wrist Left	17	Hip Right
8	Hand Left	18	Knee Right
9	Shoulder Right	19	Ankle Right
10	Elbow Right	20	Foot Right

3. PROPOSED APPROACH

This section gives a description of the proposed approach, explaining the processing of the trajectories, covariance calculations, and how the dissimilarity measures are computed to identify the subject as one of the training subjects.

3.1 Trajectory processing

The trajectories recorded using the depth camera give us the information on how each of the skeletal points are moving in the 3D space. The trajectory for each point in the skeleton, $n \in [1, 20]$ gives its position $P^n_{s,w,t} = (x^n_{s,w,t}, y^n_{s,w,t}, z^n_{s,w,t})$ in the 3D space which is a function of time (or frame number) $t \in [1, T]$ for the s^{th} subject's w^{th} walk sequence. T depends on the length of the walk sequence and our approach is independent of the duration of the walk sequence. The range of values of x, y and z depend on the scaling factors configured in the SDK and these should be normalized. The size (x and y dimensions) of the skeleton on the image plane depends on the subject's position in the camera's field of view, since the image is simply a projection of the objects in front of the camera. The skeleton size grows as the subject moves towards the camera. In order to make the identification process independent of the position of the subject in front of the camera, the skeleton size should be normalized (only the x and y dimensions, this does not apply to the depth value and the increase or decrease in depth is due to the subject moving away or towards the camera which is expected) in the time domain. The size of the whole skeleton is scaled such that the euclidean distance between the shoulder centre and the hip centre is fixed at 100 pixels through out the walk sequence. The distance between these two points are least affected when the subject walks. This makes the distance between them most suitable for use as the yard-stick for normalization. Thus we overcome the limitation of [10] that the subject has to walk parallel to the camera plane to get an almost constant skeleton height.

Since our approach tries to model the gait by analysing how the skeletal points are moving with respect to each other, it is sufficient to know the relative movement of the skeletal points with respect to a fixed point. In our approach we have chosen hip centre as the fixed point and the movement of all other points are measured relative to it. Equations (1) and (2) are used to compute at each frame, the scale factor $S_{s,w,t}$ and the relative position $\hat{P}^n_t$ of each skeletal point respectively. (It should be noted that scaling is done only for the x and y dimensions, and the depth value is not used to compute the scale factor).

$$S_{s,w,t} = \frac{dist((x^2_{s,w,t}, y^2_{s,w,t}), (x^3_{s,w,t}, y^3_{s,w,t}))}{100} \quad (1)$$

$$\hat{P}^n_{s,w,t} = (\frac{x^n_{s,w,t} - x^1_{s,w,t}}{S_{s,w,t}}, \frac{y^n_{s,w,t} - y^1_{s,w,t}}{S_{s,w,t}}, z^n_{s,w,t} - z^1_{s,w,t}) = (\hat{x}^n_{s,w,t}, \hat{y}^n_{s,w,t}, \hat{z}^n_{s,w,t}) \quad (2)$$

where $dist(A, B)$ is the euclidean distance between points A and B.

The trajectories captured using the SDI are noisy and this can be attributed to the sensor, measurement set-up and the properties of the object surface [12]. To remove the noise and achieve higher identification rates, the trajectories describing the relative movement of the points with respect to the hip centre are subjected to low pass filtering. Filtering is done by transforming the trajectories to the DCT domain and reconstructing it only with 50% of the low frequency coefficients.

3.2 Covariance and Covariance Dissimilarity

Covariance matrices and eigenvalues have been widely used for feature matching in the recognition and detection tasks such as in [13] and [14]. In [13] the authors have used covariance matrices as region descriptors. A dissimilarity measure as proposed in [15] is used for further processing to achieve detection. In our work, we apply the covariance and its dissimilarity measure concept on the skeletal trajectories.

With the trajectories of the skeletal points relative to the hip centre, the 19 skeletal points excluding the hip centre are sufficient for further analysis. For each of these skeletal points the trajectories are concatenated in the order x, y and z, which would act as a signal of 19 dimensions $\mathbf{D_{s,w}}$ for the walk sequence w of subject s.

$$\mathbf{D_{s,w}} = \left[\mathbf{G^1_{s,w}}, \mathbf{G^2_{s,w}}, ..., \mathbf{G^{19}_{s,w}}\right] \quad (3)$$

$$\mathbf{G^n_{s,w}} = \left[\hat{x}^n_{s,w,1}, ..., \hat{x}^n_{s,w,T}, \hat{y}^n_{s,w,1}, ..., \hat{y}^n_{s,w,T}, \hat{z}^n_{s,w,1}, ..., \hat{z}^n_{s,w,T}\right]'$$

The covariance matrix can model this signal by capturing the relative movement of the skeletal points which is our objective. Using the covariance matrix is a compact way of representing the walk sequence because of its low dimensionality. Unlike [9] and [10] using covariance descriptor makes

our method independent of the length of the walking sequence. It eliminates the need to segment and synchronize the walk cycles. There is also no restriction on the number of walk cycles captured per sequence for training or testing. Further the covariance is not affected by the starting position of the subject or the speed at which the subject walks. This enables us to compare walk sequences under varying conditions. Covariance matrices calculated for each walk from the training dataset of each subject forms our gait model. The covariance of these are calculated as,

$$\mathbf{C_{s,w}} = \frac{\mathbf{1}}{\mathbf{N-1}} \sum_{\mathbf{t=1}}^{\mathbf{T}} (\mathbf{d_{s,w,t}} - \mu)(\mathbf{d_{s,w,t}} - \mu)' \qquad (4)$$

where, $t \in [1, T]$ and $\mathbf{d_{s,w,t}}$ is a 19 dimensional feature point (row vector of $\mathbf{D_{s,w}}$) at time instant t, $N = 19$ and μ is the mean of the samples. For a given test walk sequence, the trajectories are processed in the same way as described earlier in section 3.1 and its covariance matrix is computed.

A dissimilarity measure is computed between the test covariance matrix and the model covariance matrices,

$$\delta(\mathbf{C_{test}}, \mathbf{C_{s,w}}) = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{ln^2} \lambda_\mathbf{i}(\mathbf{C_{test}}, \mathbf{C_{s,w}})} \qquad (5)$$

where $\lambda_i(\mathbf{C_{test}}, \mathbf{C_{s,w}})$ are the generalized eigenvalues of $\mathbf{C_{test}}$ and $\mathbf{C_{s,w}}$ that satisfy $\mathbf{C_{test}x} = \lambda \mathbf{C_{s,w}x}$, and $\mathbf{x}$ is the corresponding generalized right eigenvector. The dissimilarity measure between two symmetric positive definite matrices $\mathbf{C_a}$ and $\mathbf{C_b}$ satisfies the following:

1. $\delta(\mathbf{C_a}, \mathbf{C_b}) \geq \mathbf{0}$ and $\delta(\mathbf{C_a}, \mathbf{C_b}) = \mathbf{0}$ only if $\mathbf{C_a} = \mathbf{C_b}$.
2. $\delta(\mathbf{C_a}, \mathbf{C_b}) = \delta(\mathbf{C_b}, \mathbf{C_a})$.
3. $\delta(\mathbf{C_a}, \mathbf{C_b}) + \delta(\mathbf{C_b}, \mathbf{C_c}) \geq \delta(\mathbf{C_a}, \mathbf{C_c})$.

The subject of the test walk sequence is identified by comparing the test covariance matrix with those in the training set. The label of the least dissimilar training covariance matrix is identified as the subject. The recognition results obtained are identical when the training and testing sequences are swapped because of the symmetry property of the dissimilarity measure.

4. TESTING AND RESULTS

4.1 Database

For the experiment, we created a database of 20 subjects using the *Kinect* depth sensor, recorded at approximately 30 frames per second. 20 different people were asked to walk in front of the camera 10 times, which gave a total of 200 walk sequences in two scenarios. For the first 7 walk sequences of each subject, the subject is asked to walk straight towards the camera. For the remaining three sequences, the subject is asked to walk along an arbitrary straight line and the camera is manually panned to capture the subject's movement. The gait model is created using a certain number of walk sequences depending on the testing combinations listed in Table 2 and the remaining walk sequences are used for testing.

4.2 Results

The proposed approach was tested by choosing various combinations of walk sequences for training and testing. The combinations and the recognition rates achieved for these are summarized in Table 2. The table shows that we have achieved recognition rates above 90% for all combinations. Testing combination 1 shows that our approach gives very high recognition rates (97.5%) when the test walk sequences and the training walk sequences both are from the first scenario. High recognition rates were achieved also when the walk sequences (8,9 and 10) from the second scenario were added to the testing combinations, which shows that the proposed method is robust to variations such as the walking direction and the camera view angle. This is observed from the testing combinations 2 and 3. Figure 4 shows the confusion matrix and Figure 5 shows a plot of the recognition rate at 4 ranks for different test combinations. Rank plot indicates the presence of actual subject in the top k ranks. Table 3 compares the result of the proposed method with that of [9] and [10] which work on the same type of data obtained using the *Kinect* sensor.

In Table 4, the recognition rates for different groupings of the skeleton parts (spine, left arm, right arm, left leg and right leg) are shown for test combinations 1 and 4. When the parts of the skeleton are used individually the recognition performance is low as compared to when they are used together. Using arms alone, legs alone or a combination of those gives a recognition rate of 75% to 93.75%. Introduction of spine along with one another part gives performance of 75% to 97.5%. This shows that at least two parts of the skeleton have to be considered together to achieve a notable performance. It is also in accordance with our intuition that relative movement information of the body parts significantly represents the gait of a subject. The combination of spine, arms and legs gives a better performance than using the legs alone. This shows that the upper-body data is as important as the lower body data and can improve the results significantly.

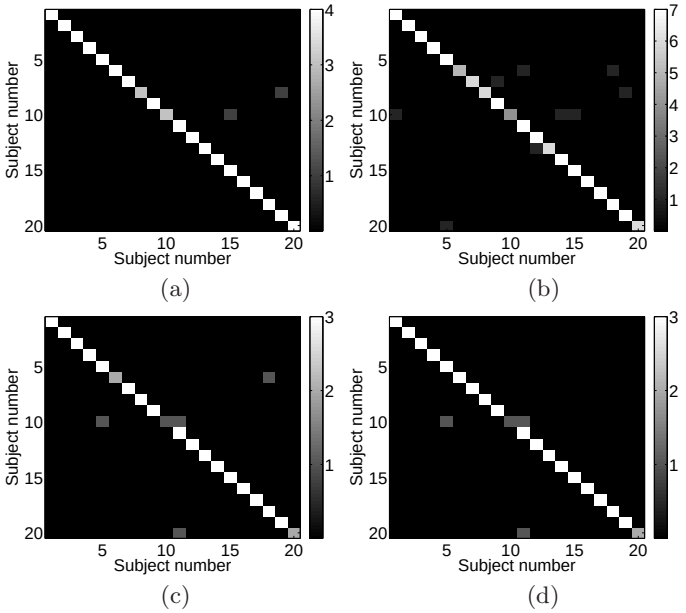

Figure 4: Confusion matrix for test combination: (a) 1 (b) 2 (c) 3 (d) 4

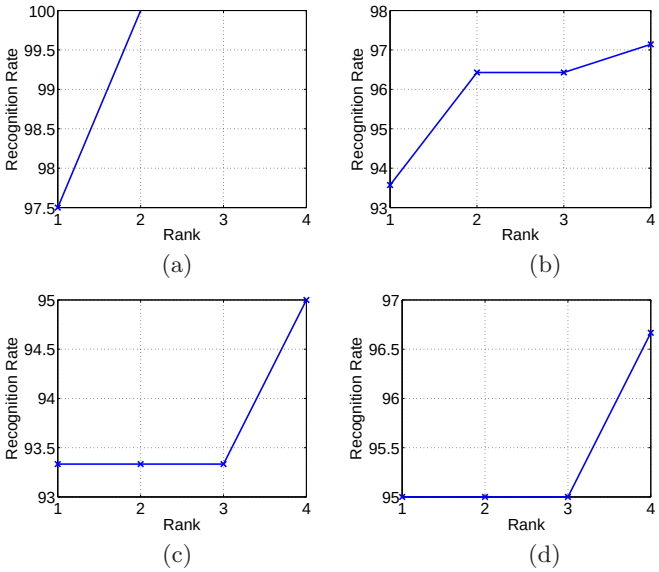

Figure 5: Recognition rate v/s rank for test combination: (a) 1 (b) 2 (c) 3 (d) 4

Table 2: Recognition rates for various testing combinations

Test Combination	Training Walks	Test Walks	Recognition Rate(%)
1	1,2,3	4,5,6,7	97.5
2	1,2,3	4,5,6,7,8,9,10	93.57
3	1,2,3,4,5	8,9,10	93.33
4	1,2,3,4,5,6,7	8,9,10	95

5. CONCLUSION AND FUTURE WORK

In this paper, a novel gait recognition approach was proposed based on skeleton-point trajectories obtained from a *Kinect* depth camera. Depth sensor provides a new dimension of features for gait analysis. It provides a more robust 3-D skeleton information since depth information provides an accurate silhouette sequence. The proposed approach achieves recognition performance of above 90% in an unrestricted indoor ambience. The robustness of the proposed approach is quantified for fixed and moving camera scenarios. We are in the process of expanding the dataset with more number of subjects and walking scenarios, with challenges like occlusion.

6. ACKNOWLEDGMENTS

This work was partly supported by DRDO-CARS project 'Compressed Domain Human Activity Analysis' (Ref: CAIR-CARS-25).

7. REFERENCES

[1] N.V. Boulgouris, D. Hatzinakos, and K.N. Plataniotis. Gait recognition: a challenging signal processing technology for biometric identification. *IEEE Signal Processing Magazine*, 22(6):78–90, 2005.

[2] G. Johansson. Visual motion perception. *Scientific American*, pages 76–88, 1975.

Table 3: Recognition Rate comparison

Method	Proposed	[9]	[10]
Recognition Rate	97.5%	43.6%	91.0%

Table 4: Recognition rates for various groups of points

Skeletal point groups	Recognition rate(%)	
	Test comb. 1	Test comb. 4
spine	56.25	43.33
left arm	58.75	48.33
right arm	68.75	50
left leg	35	51.67
right leg	46.25	41.67
left arm,left leg	85	75
right arm,right leg	93.75	85
left arm,right arm	90	76.67
left leg,right leg	78.75	81.67
left arm,right leg	87.5	81.67
right arm,left leg	95	85
spine,left arm	88.75	78.33
spine,right arm	97.5	75
spine,left leg	83.75	75
spine,right leg	83.75	66.67
spine,left arm,left leg	92.5	75
spine,right arm,right leg	98.75	85
spine,left arm,right arm	93.75	86.67
spine,left leg,right leg	93.75	85
spine,left arm,right leg	97.5	86.67
spine,right arm,left leg	100	93.33

[3] S.A. Niyogi and E.H. Adelson. Analyzing and recognizing walking figures in XYT. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 469–474. IEEE, 1994.

[4] A. Johnson and A. Bobick. A multi-view method for gait recognition using static body parameters. In *Proceedings of Conference in Audio and Video based Biometric Person Authentication*, pages 301–311, 2001.

[5] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern Recognition Letters*, 17(2):155–162, 1996.

[6] Toby H.W. Lam, Raymond S.T. Lee, and David Zhang. Human gait recognition by the fusion of motion and static spatio-temporal templates. *Pattern Recognition*, 40(9):2563–2573, 2007.

[7] Changhong Chen, Jimin Liang, Heng Zhao, Haihong Hu, and Jie Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.

[8] Sabesan Sivapalan, Daniel Chen, Simon Denman, Sridha Sridharan, and Clinton Fookes. Gait energy volumes and frontal gait recognition using depth images. In *Proceedings of International Joint Conference on Biometrics*, pages 1–6. IEEE, 2011.

[9] Adrian Ball, David Rye, Fabio Ramos, and Mari Velonaki. Unsupervised clustering of people from skeleton data. In *HRI '12 Proceedings of the seventh*

annual ACM/IEEE international conference on Human-Robot Interaction, pages 225–226, 2012.

[10] Johannes Preis, Moritz Kessel, Martin Werner, and Claudia Linnhoff-Popien. Gait recognition with kinect. In *1st International Workshop on Kinect in Pervasive Computing*, 2012.

[11] Microsoft SDK for Kinect:. http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk/.

[12] K. Khoshelham. Accuracy analysis of kinect depth data. In *ISPRS Workshop Laser Scanning*, 2011.

[13] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, volume 3952, pages 589–600, 2006.

[14] Chengjun Liu and Harry Wechsler. A shape- and texture-based enhanced fisher classifier for face recognition. *IEEE Transactions on Image Processing*, 10(4):598–608, 2009.

[15] Wolfgang Forstner and Boudewijn Moonen. A metric for covariance matrices. Technical report, Dept. of Geodesy and Geoinformatics, Stuttgart University, Germany, 1999.