Kinect based Frontal Gait Recognition using skeleton and depth derived features

Manasa Gowri Hebbur Sheshadri, Manish Okade
Department of Electronics and Communication Engineering,
National Institute of Technology Rourkela, India
manasagowri16@gmail.com, okadem@nitrkl.ac.in

***Abstract*—Recognizing humans through gait has been an emanant biometric technology in the recent years owing to the fact that it is unobtrusive since it does not require a subject's cooperation. This paper investigates Kinect based gait recognition of human subjects for surveillance applications especially in narrow corridor and airport scenarios where only the frontal views are available. Two features namely skeleton size feature and projectile motion feature extracted from skeleton data and one feature derived by segmenting the depth data using superpixels followed by SURF descriptor extraction are utilized in a hierarchical framework to obtain the closest matching subject for recognition purposes. The proposed method provides considerable increase in the recognition accuracy and recognition rank in comparison to state-of-the-art gait recognition approaches.**

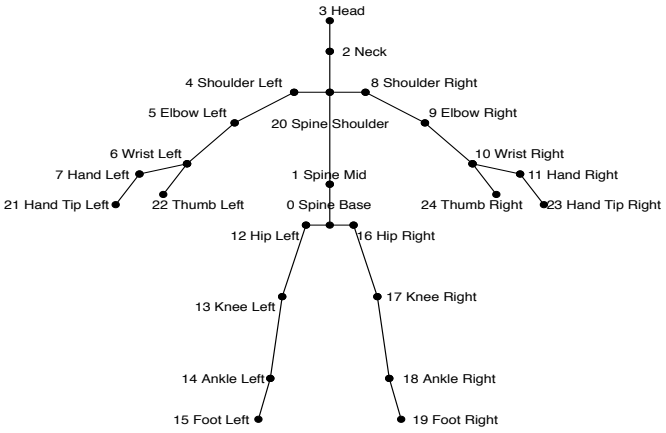
***Index Terms*—Human Gait, Kinect camera, Skeleton data, Depth data, Frontal Gait, kNN Classifier.**

I. Introduction

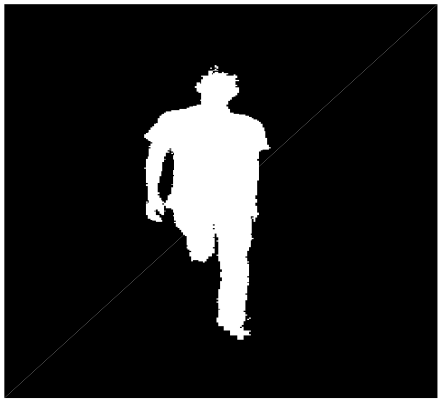
Biometric authentication has been the most sought-after technology for security systems in the recent years for numerous reasons. Various techniques have been proposed for biometric authentication including fingerprint, face, iris, gait, etc., among which gait which refers to the walking style of a person has a lot of benefits. It is non invasive, can be characterized using low resolution videos and difficult to be masked. Significant research work has been done in gait based human identification in the past decade. Most of the early works on vision based gait recognition considers the availability of fronto parallel (side) view, since this view captures maximum amount of information about a person's gait. Techniques such as Gait Energy Image [1], Pose Energy Image [2] wherein binary silhouettes obtained from the side view have been utilized for recognition purposes are good examples of fronto parallel methods. In addition, motion based descriptors like SURF [3], [4] have also been used for gait recognition considering the fronto parallel view. However, most of the surveillance scenarios like airports, railway stations and narrow corridors, frontal view of a person is more likely to be captured effectively than the fronto parallel view. Hence there is a need for gait recognition techniques from frontal view.

RGB cameras may prove to be incapable of capturing gait effectively in such scenarios, thus posing the need for depth views which inturn necessitate the need for depth sensing cameras like Kinect [5] which provide skeleton and depth related information in 3 dimensions. Utilizing only skeleton information from the frontal view [6] although is computationally efficient, yet it suffers from lower recognition accuracy. On the other hand, utilizing only depth information [7], [8] although gives good recognition accuracy, yet it suffers from being computationally complex. A good tradeoff was attempted in [9] where both skeleton and depth information were used that achieved good results.

978-1-7281-5120-5/20/$31.00 © 2020 IEEE



(a)



(b)

Fig. 1: (a) Skeleton Joint structure provided by Kinect, (b) Depth Image provided by Kinect

However, this method suffered from a few disadvantages. Firstly, it assumes the presence of two cameras and thus implying that videos from two different views should be available for detection. Secondly, the depth based feature obtained suffers from lack of dynamic information since the silhouettes have been binarized and averaged. The method proposed in this paper overcomes these limitations by using a single camera instead of two thus reducing the camera overhead and also by using SURF (Speeded Up Robust Feature) features which preserve the depth information.

II. Novel Contribution

The proposed method uses the skeleton information provided by Kinect as shown in Fig 1(a) to estimate the angle made by the knee joints and the hip joints as a person walks in front of the camera. Presence of 3D skeleton information ensures that the angle variation can be captured from the required plane of view, i,e., the knee joint angle variation can be best captured in sagittal(YZ) plane whereas the hip joint angle variation can be best captured in transversial(XZ) plane. Thus, the presence of 3D joint information ensures that the corresponding angles can be captured effectively even when the camera is placed in the front view. Furthermore, the use of SURF descriptors on depth image as shown in Fig 1(b) ensure the effective capture of dynamic information about the variations in the subject's body parts even from the front view. Thus, the proposed method captures gait information in a more efficient manner from the front view thereby ensuring higher recognition rates. The extracted features are such that for a specific person's walk the features are grouped together in the feature space and thus a simple distance based classifier such as k-nearest neighbor (kNN) will be efficient in classifying the person's gait.

III. Proposed method

The block diagram of the proposed method is shown in Fig 2. The proposed method consists of extracting two novel features from skeleton data and one novel feature from depth data. The three features extracted are combined using a hierarchical classification strategy, as it reduces the number of gallery sequences to be compared at every level of hierarchy. Since skeleton data is faster to process in comparison to depth data, it has been used in the initial two levels of the hierarchical scheme.

Two skeleton features namely, skeleton size feature and projectile motion feature derived from skeleton information (Fig. 1(a)) are utilized at the first two levels of hierarchy since they provide information about the 25 joint positions of a test subject. Hence by eliminating vastly dissimilar subjects in the initial levels of hierarchy using these skeleton features, the search space is reduced which leads to lower computational complexity. In the third level of hierarchy, the depth image obtained from Kinect (Fig. 1(b)) is segmented using superpixel segmentation followed by extracting SURF features from the segmented image. This gives the detailed information about the subject and helps in achieving higher recognition accuracy. All the features are extracted for one gait period of a subject. The gait period estimation is carried out as proposed in [9] wherein the absolute depth difference between the limbs is utilized in calculating the gait period and the details are omitted here due to space constraints.

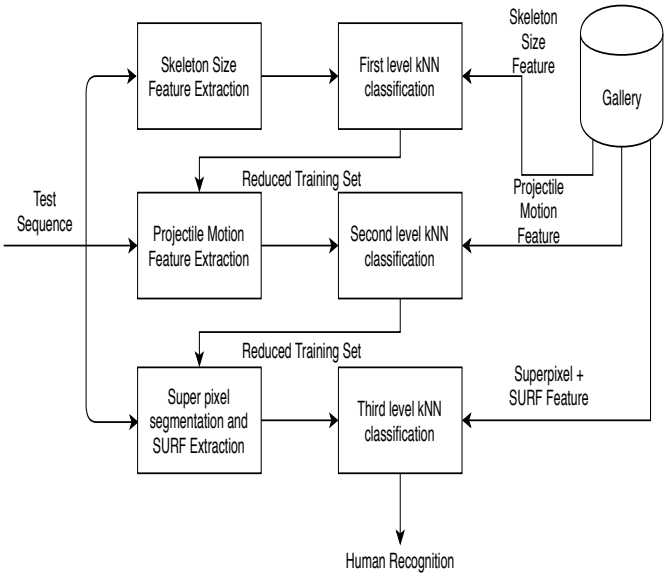


Fig. 2: Block diagram of the proposed method

A. Skeleton Size Feature Extraction

The height and width of a particular subject as seen from the front view can be used as filters to restrict the search during biometric identification. Suppose 'N' frames $F_1, F_2, ..., F_N$ constitute a gait period, the height H^n [9] for the n^{th} frame is computed as given below;

$$H^n = d^n_{3,2} + d^n_{2,20} + d^n_{20,1} + d^n_{1,0} + \max(d^n_{12,13} + d^n_{13,14}, d^n_{16,17} + d^n_{17,18}) \quad (1)$$

where $d^n_{i,j}$ is the distance between i^{th} and j^{th} joint positions as shown in Fig 1(a) in the n^{th} frame. Width of the subject is taken to be the separation between the left and the right elbow along the X axis and is given by;

$$W^n = |x^e_r - x^e_l| \quad (2)$$

where x^e_r and x^e_l are the x positions of the left and right elbow joints respectively. The average of the height and width values for all the frames in a gait cycle are computed as $H = mean(H^n)$ and $W = mean(W^n)$ and concatenated to form the first feature vector ($\mathcal{V}_1$) as given below;

$$\mathcal{V}_1 = [H, W] \quad (3)$$

B. Projectile Motion Feature Extraction

This feature describes the trajectories of the angles made by various joints while a subject is walking and is referred as the Projectile Motion Feature. The joint angles from the lower

body joints are vividly distinguishable to the human eye and hence we use these joint angles as feature. Trajectories of four different angles (i.e., their values in 30 frames of a gait period) are computed for the left and right joints of the body as follows;

Fig. 3: Skeleton transformation

1) Transversial Plane Hip Joint Angle Estimation: : The angle made by the left and the right hip joints w.r.t., the hip center in the transversial plane (i.e., XZ plane of the Kinect coordinate system) as shown in Fig 3 is computed. Since the skeleton is transformed to coincide hip center with the origin of the Kinect coordinate system, the hip centre joint position is (0,0,0). Let (x_l^h, y_l^h, z_l^h) denote the position of the left hip joint and (x_r^h, y_r^h, z_r^h) denote the position of the right hip joint for the n^{th} frame, then the angles between left and right hip joint and the hip center is estimated using;

$$\theta_{dh}{}^n = \tan^{-1}(\tfrac{z_d^h}{x_d^h}) \tag{4}$$

where $d = [l, r]$ for the left and right joints respectively.

2) Sagittal Plane Knee Joint Angle Estimation: : The angles between the left and right knee, left and right ankle and hip center at the knee joint in the sagittal plane (i.e., YZ plane of the Kinect coordinate system) is computed. Let (x_l^k, y_l^k, z_l^k) and (x_r^k, y_r^k, z_r^k) denote the coordinates of the left and right knee joints respectively and (x_l^a, y_l^a, z_l^a) and (x_r^a, y_r^a, z_r^a) denote the coordinates of the left and right ankle joints respectively, then the slopes of the lines formed by the hip center and knee (m_d^1), knee and ankle (m_d^2) are as given below;

$$m_d^1 = \frac{y_d^k}{z_d^k} \quad \text{(5a)} \qquad m_d^2 = \frac{y_d^a - y_d^k}{z_d^a - z_d^k} \quad \text{(5b)}$$

The required angle is now estimated using;

$$\theta_{da}^n = \tan^{-1}\left(\left|\tfrac{m_d^2 - m_d^1}{1 + m_d^1 m_d^2}\right|\right) \tag{6}$$

where $d = [l, r]$ for left and right joints respectively and n is the frame considered.

Finally, the trajectories of the four angles are concatenated over all the frames of a gait period to form the Projectile Motion Feature ($\mathcal{V}_2$) as given below;

$$\mathcal{V}_2 = [\theta_{lh}, \theta_{rh}, \theta_{la}, \theta_{ra}] \tag{7}$$

C. Superpixel Segmentation and SURF Extraction

The depth image obtained from Kinect SDK (Fig. 1(b)) is now used to obtain the final set of features. The depth image is a grayscale image whose intensity at a pixel depends on the depth of the corresponding pixel w.r.t. the Kinect sensor. From all the frames available in a gait cycle, 10 key frames are used for further processing. For the n^{th} key frame F_n, the depth image is subjected to preprocessing to remove noise. Superpixel segmentation using the Simple Linear Iterative Clustering (SLIC) algorithm [10] is then applied and the superpixel image thus obtained is used to extract Speeded Up Robust Features (SURF) [11]. The features thus obtained for all the 10 key frames are then concatenated to form the SURF based feature $\mathcal{V}_3$. 10 key frames were chosen amongst all frames of a subject since it was observed during experimental simulation that they were enough to characterize the walking pattern of the subject and an increase in the number of key frames beyond ten did not affect the accuracy estimation significantly.

IV. Experimental Results

The proposed method is used for human identification by comparing the gait of a test subject against a large gallery set. In the present study the training gallery set size is 60 so that comparative analysis with existing methods is uniform. The expermentation is carried out using MATLAB R2018a on a 1.80GHz Intel Core i7 processor having 12GB RAM. The results reported in this paper are reproducible and the MATLAB source code is available at https://sites.google.com/site/manishokade/publications.

A. Dataset Description

Fig. 4: Camera setup for data collection using two Kinect cameras for the front and back views

Over the past few years, availability of Kinect camera has simplified the job of depth based feature extraction, since Kinect SDK directly gives the skeleton and depth information. Hence, we have constructed a new dataset using Kinect camera which will be made public for fellow researchers after the review cycle is completed. The setup and recording procedure for the captured dataset is explained below.

A total of 60 distinct subjects were used to capture the dataset. The recording setup for capturing the dataset is as shown in Fig 4. Each Kinect is placed at a height of 1.5 metres from the ground using a height adjustable tripod. The subject is made to walk in between the two cameras in their viewing range. The skeleton information of Kinect and the depth information provided by Kinect as shown in Fig 1 are captured for each frame. 10 sequences are captured for each subject out of which the sequences used for training and testing for different test combinations is shown in Table I. The numbers present in the training and test walks columns represent the sequence at which the walks were captured from 1 to 10.

B. Result Analysis

The number of key frames analyzed in a gait cycle for a given subject is taken to be a tunable parameter. Performance analysis is carried out by varying this parameter $m \in (5, 10, 20, 30)$ for the test combination T4. As observed from Table II for 10 key frames, both accuracy and computational complexity of the SURF feature reach optimal tradeoff values.

TABLE I: Test Combinations

Test Combination	Training Walks	Test Walks
T1	1,2	9,10
T2	1,2,3,4	9,10
T3	1,2,3,4,5,6	9,10
T4	1,2,3,4,5,6,7,8	9,10

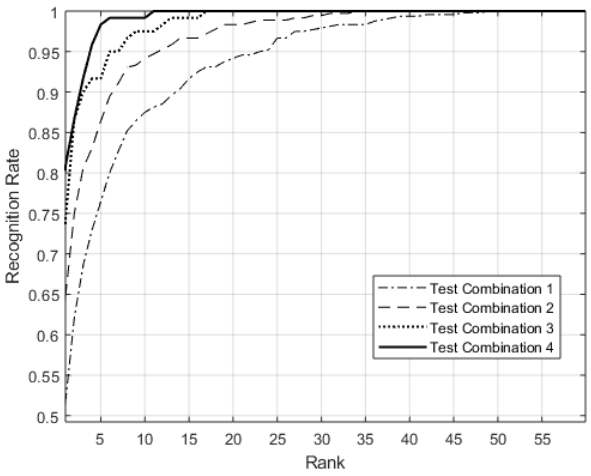
TABLE II: Performance analysis of SURF feature for different number of key frames

Number of key frames	Accuracy (%)	Time (secs)
5	74.17	231
10	**77.50**	**460**
20	79.17	990
30	77.50	1387

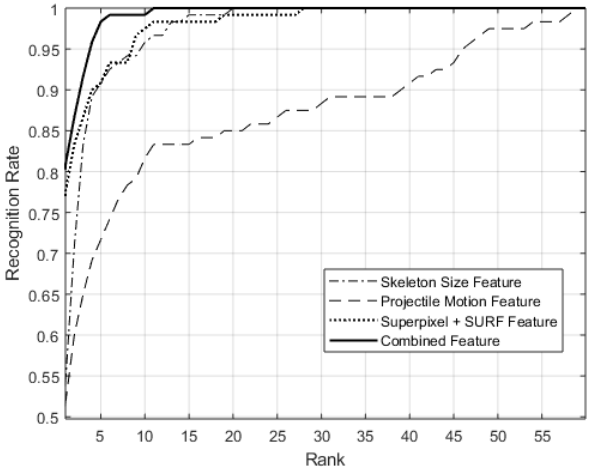
The recognition accuracy for each of the proposed individual features and the combined feature for different test combinations is shown in Table III. These results show that as the number of training samples is increased the recognition accuracy improves. This is expected since increase in training samples implies increase in the data for comparison and hence improvement in accuracy. Additionally, it can be observed from Table III that the combined feature has greater accuracy in comparison to utilizing individual features independently.

TABLE III: Comparison of performance of proposed method for different test combinations

	T1		T2		T3		T4	
Gait Feature	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
Skeleton Size Feature	47.50	0.05	45.83	0.10	50.00	0.13	54.17	0.18
Projectile Motion Feature	36.67	0.07	45.00	0.12	44.17	0.17	51.67	0.26
Superpixel + SURF Feature	53.33	108.00	65.83	221.00	72.50	340.00	77.50	460.00
Combined Feature	53.33	85.00	67.50	134.00	74.17	217.00	**80.83**	**195.00**



(a)



(b)

Fig. 5: CMC curve showing recognition rate versus rank of the proposed method for (a) different test combinations, (b) the derived features

Finally, we analyze the Cumulative Match Characteristic (CMC) curves for different test combinations. CMC curve is a precision curve that provides recognition precision for each rank. In many real life scenarios, it is sufficient if we can predict that the correct class falls within top r% of the classes predicted by the algorithm. In this regard, we present the recognition accuracy using CMC curves for the proposed method in Fig 5(a) for different test combinations given in Table I. The curve indicates that the recognition accuracy improves for all ranks as the number of training sequences increases which is on expected lines.

The CMC curve is also analyzed in Fig 5(b) for each of the individual features along with the combined feature. The curve shows that the recognition accuracy of the combined feature is greater than each of the individual features and also the accuracy improves exponentially with rank and reaches 100% at rank 11. This is due to the fact that at each level

of hierarchy vastly dissimilar elements are eliminated from comparison thereby improving the accuracy and at the same time reducing the complexity of using only the depth based features.

TABLE IV: Comparative analysis with existing state-of-the-art methods

Gait Recognition Technique	Accuracy (%)	Time (secs)
Hierarchical Approach [9]	68.33	1.04
Covariance Approach [6]	60.00	3.63
Proposed Method	80.83	195.00

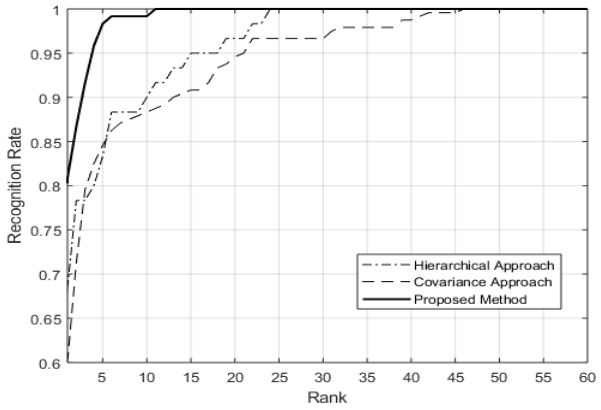


Fig. 6: CMC curve showing recognition rate versus rank for different gait recognition techniques

Comparative analysis of the proposed method with the existing methods is shown in Table IV and Fig. 6. As observed the proposed method is superior to existing methods both in terms of recognition accuracy as well as in terms of recognition rank indicating that the combined feature is a good alternative in terms of performance as compared to existing shape and motion features. The processing time of the proposed features can be reduced by using more efficient coding techniques and also using better processors for computing.

V. Conclusions

In this paper, three different features were utilized for gait recognition using frontal view gait sequences captured using Kinect camera. Two features were extracted from skeleton data and used in the first two levels of the hierarchical scheme while the third feature was derived out of the depth data by carrying out segmentation of the depth image and utilizing the last stage of hierarchy. The three features were combined using a hierarchical classification strategy using a kNN classifier. The results obtained indicate that the features combined together perform better than two of state-of-the-art methods for gait recognition by providing an accuracy of 80.83% and an accuracy of 100% for a rank as small as 11. All the features are extracted using only front view sequences provided by Kinect. Hence, the proposed method is suitable in narrow corridor like scenarios where only front view gait sequences are available.

References

[1] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[2] A. Roy, S. Sural, and J. Mukherjee, "Gait recognition using pose kinematics and pose energy image," *Signal Processing*, vol. 92, no. 3, pp. 780–792, 2012.

[3] N. N. Htwe and N. War, "Human identification based biometric gait features using msrc," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 2, no. 5, 2013.

[4] M. H. Khan, F. Li, M. S. Farid, and M. Grzegorzek, "Gait recognition using motion trajectory analysis," in *International Conference on Computer Recognition Systems*. Springer, 2017, pp. 73–82.

[5] S. D. Mowbray and M. S. Nixon, "Automatic gait recognition via fourier descriptors of deformable objects," in *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 566–573.

[6] M. Kumar and R. V. Babu, "Human gait recognition using depth camera: a covariance based approach," in *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*. ACM, 2012, p. 20.

[7] P. Chattopadhyay, A. Roy, S. Sural, and J. Mukhopadhyay, "Pose depth volume extraction from rgb-d streams for frontal gait recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 53–63, 2014.

[8] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Gait energy volumes and frontal gait recognition using depth images," in *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–6.

[9] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal gait recognition from incomplete sequences using rgb-d camera," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1843–1856, 2014.

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.