Exploiting Pose Information for Gait Recognition from Depth Streams

Pratik Chattopadhyay[1], Shamik Sural[1], and Jayanta Mukherjee[2]

[1] School of Information Technology, IIT Kharagpur, India
[2] Department of Computer Science & Engineering, IIT Kharagpur, India
pratikc@sit.iitkgp.ernet.in,shamik@sit.iitkgp.ernet.in,
jay@cse.iitkgp.ernet.in

Abstract. A key-pose based gait recognition approach is proposed that utilizes the depth streams from Kinect. Narrow corridor-like places, such as the entry/ exit points of a security zone, are best suited for its application. Alignment of frontal silhouette sequences is done using coordinate system transformation, followed by a three dimensional voxel volume construction, from which an equivalent fronto-parallel silhouette is generated. A set of fronto-parallel view silhouettes is, henceforth, utilized in deriving a number of key poses. Next, correspondences between the frames of an input sequence and the set of derived key poses are determined using a sequence alignment algorithm. Finally, a gait feature is constructed from each key pose taking into account only those pixels that undergo significant position variation with respect to the silhouette center. Extensive evaluation on a test dataset demonstrates the potential applicability of the proposed method in real-life scenarios.

Keywords: Gait recognition, depth camera, key pose, incomplete cycle sequences, variance image

1 Introduction

Constant monitoring of subjects and identification of suspects are essential activities for providing public security inside crowded security zones. Human recognition using biometric identification mechanisms like finger print detection and iris scan cannot be employed in these congested places, since these methods require close interaction with subjects. Face recognition is also not convenient in such a scenario because highly detailed texture information in face images might be missing in surveillance videos, as they are usually captured from a distance. Gait is the only biometric which can possibly be applied to identify suspects in these congested security areas. Till date, a number of computer vision based gait recognition algorithms, corresponding to both the fronto-parallel [1–4] as well as the frontal views [5, 6] have been developed, each of which has been shown to work effectively with low resolution gait video sequences.

Although it is known that gait video from the fronto-parallel view captures significant information about an individual's gait [7], in real-life, it is possible to encounter situations where surveillance needs to be carried out in narrow corridor-like places, such as the entry/ exit points of security zones. Due to the constricted field of view of a surveillance camera placed within a narrow region, it might not be able to capture sufficient number of frames of a walking sequence required for analyzing the gait characteristics of the concerned subject. On the other hand, a relatively higher percentage of frames can be recorded if the camera is positioned in a way so as to capture walking videos from the front view.

However, a drawback associated with frontal gait recognition is that, information about the important fronto-parallel component of gait [7] cannot be obtained from the silhouettes captured by an RGB camera from the frontal view. It appears that a substantial fronto-parallel information of the gait of a silhouette can be extracted even from the frontal view if the knowledge about its three dimensional pose is available. Gait videos collected using a depth camera like Microsoft Kinect [8] seems to be beneficial in this aspect. In this paper, we propose to carry out gait recognition using Kinect as a surveillance camera. The gait recognition scenario considered here is shown in Fig. 1. With reference to the figure, the Kinect (C), used as a surveillance camera is installed at a certain height above a narrow pathway. As a subject walks through the pathway along the direction shown in the figure, C captures the depth information of the gait of the subject from the front view. This depth information is, henceforth, utilized in deriving an effective gait feature which preserves significant fronto-parallel information. Experimental results on an extensive dataset proves the efficacy of the proposed method in real-life scenarios.

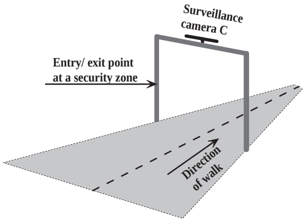


Fig. 1. Camera setup for gait recognition in a narrow security zone

The rest of the paper is organized as follows. Section 2 provides a brief background study on the recent research trend in frontal gait recognition using depth information. Construction of the gait feature and human recognition using the derived feature is explained in Section 3. A detailed description of the dataset along with experimental results is presented in Section 4. Section 5 concludes the paper and points out future scope of work.

2 Literature Survey

Early gait recognition approaches [1–4] use gait videos captured by RGB cameras and focus mostly on the fronto-parallel view of gait. Development of depth cameras like Kinect [8] has resulted in shifting of focus towards frontal gait recognition using depth information [5, 6, 9–11]. Among the existing depth cameras, Kinect [8], developed by Microsoft, has gained significant popularity in human tracking based research, primarily because of the useful human detection and skeleton tracking application [12] provided by the Kinect SDK. In the recent past, Kinect has been extensively used in deriving interesting frontal gait features that make use of both the skeleton streams [9–11] and the depth streams [5, 6] obtained from its SDK. Each of these techniques has shown promising results, but a few assumptions inherent in these methods limit their applicability in real-life scenarios.

In [9], a gait feature using the skeleton joint coordinates is proposed by Kumar et al., in which the covariance of each joint trajectory over a complete gait cycle is used in recognition. Milovanovi et al., in [10], describe a recognition scheme where skeleton data of a gait cycle is mapped from the spatial domain to the spatio-temporal domain and content-based image retrieval techniques are applied for feature construction. Both these methods have been shown to work satisfactorily in the presence of complete gait cycles, but their performance in the absence of full cycle information is unclear. Chattopadhyay et al. propose a frontal gait recognition approach in [11], where a complete gait cycle is divided into a fixed number of key poses derived from the skeleton structure provided by Kinect SDK. The gait cycle partitioning scheme, as proposed in this work, has been shown to outperform [9] as well as a traditional binary silhouette based gait recognition scheme, namely, the technique using Gait Energy Image (GEI) [1]. Although gait recognition methods using the skeleton streams from Kinect have significantly fast response time, their effectiveness depends on the accuracy of the skeleton joints tracked by the SDK. Also, lack of complete silhouette shape/ depth information, because of using only the skeleton streams from Kinect, might have a negative impact on the accuracy of gait recognition.

Only a few approaches use solely the depth streams from Kinect to derive frontal gait features [5, 6, 13]. Hofmann et al. describe a gait recognition procedure in [13], where gradient histograms computed from the depth frames of a fronto-parallel view gait sequence are averaged over a gait cycle. The results presented in this work show that the use of depth information helps in achieving a higher recognition rate than GEI [1]. However, since no significant depth variation occurs in the fronto-parallel view gait sequences, it is unclear if the use of depth information has significant benefits for gait recognition from this view. The work by Sivapalan et al. in [5] provides an effective means of utilizing the depth streams from Kinect in carrying out gait recognition from the frontal view. The feature proposed in [5] is termed as Gait Energy Volume (GEV). It is derived by averaging the voxel volumes constructed from the corresponding point cloud sequences captured by Kinect over an entire gait cycle. But, as also explained in [2], such averaged information lacks intrinsic kinematic details about the gait of a subject. Hence, GEV fails to perform satisfactorily, if there exist a number of corrupted/ noisy silhouettes in the sequence, or if the Kinect fails to capture at least one complete gait cycle of the walking subject.

To overcome the limitations of GEV, a pose based feature termed as Pose Depth Volume (PDV) was proposed by Chattopadhyay et al. in [6]. Here, noisy depth silhouettes are initially smoothed by registering each depth frame with the corresponding RGB frame. PDV helps in preserving the dynamic component of gait at a higher resolution than GEV because the feature is derived at the granularity of key poses. However, the expensive voxel level computation in PDV and also the requirement of at least a complete gait cycle for its satisfactory performance, make it impractical for use in real-life situations.

It appears from the algorithms proposed in [2] (PEI) and [6] (PDV) that, carrying out gait recognition at the granularity of key poses significantly enhances the efficacy of recognition. This motivates us in proposing a key pose based gait recognition approach in order to carry out recognition in the scenario considered in the present paper (refer to Section 1, Fig. 1). In contrast to the existing frontal gait recognition techniques, the proposed method effectively preserves the important gait information corresponding to the fronto-parallel view by making use of the three dimensional depth information of the silhouette points provided by Kinect. Moreover, recognition from complete gait cycle information as considered in each of the techniques given in ([1, 2, 5, 6, 9–11]) cannot be regarded as a practical solution. Such constraint on the minimum length of a gait cycle is potentially eradicated in the present paper by carrying out subject identification using only the available key poses in a given sequence. This adds a higher degree of pertinency to the proposed approach as compared to the state-of-the-art gait recognition techniques in application sites similar to Fig. 1. The main contributions of the paper can be summarized as follows:

- Development of a frontal gait recognition technique when number of training samples is few and also when unconstrained data are captured with no restriction on the minimum length of the gait cycle,
- derivation of equivalent fronto-parallel view silhouettes by utilizing the depth information of the frontal surface of silhouettes recorded by Kinect and extraction of gait features from these silhouettes, thereby, preserving important gait information, and
- extensive experimental evaluation emphasizing the effectiveness of the proposed approach.

3 Proposed Approach

As described in Section 1, we propose a key pose based frontal gait recognition approach using Kinect captured datasets. Surveillance cameras inside a security zone, are usually mounted at a certain height facing downwards. The gait recognition scenario described in Section 1 also takes into account of a similar Kinect camera setup. Hence, the point cloud of a walking subject as captured by the Kinect is inclined with respect to the Kinect coordinate system. The recognition procedure must be made invariant to the tilt angle since this angle may vary from one surveillance site to another.

3.1 Alignment of Silhouette Sequence and Construction of Voxel Volumes

Invariance to the camera tilt angle is achieved by applying a set of geometric transformation operations on each point cloud and also by aligning it with respect to a fixed coordinate system. The alignment operation is explained with the help of Fig. 2.

Coordinate System Transformation The objective of this alignment procedure is to obtain an upright silhouette point cloud that will be perpendicular to the viewing direction. To achieve this, we determine transformed point cloud coordinates with respect to a different coordinate system (say, X', Y', Z'), such that, the $X'Y'$ plane of this coordinate system is parallel to the direction of orientation of the point cloud, and the Z' axis is along a direction normal to this plane. Without loss of generality, we consider that the origins of the Kinect coordinate system and the (X', Y', Z') coordinate system coincide. As seen in Fig. 2, initially linear regression [14] of the set of object points P present in the Kinect captured point cloud is used in determining a plane $\mathcal{P}$ that passes through the origin of the Kinect coordinate system, and is closely parallel to the direction of alignment of the point cloud. Corresponding to each object point P_i with coordinates (X_i,Y_i,Z_i) (measured in the Kinect coordinate system), we determine its transformed coordinates P'_i (X'_i,Y'_i,Z'_i) with respect to the (X', Y', Z') coordinate system. Let the unit normal vector to the plane $\mathcal{P}$ constructed from the point set P be given by (nx, ny, nz) (as shown in Fig. 2). If the equation of this plane is given by $Z = AX + BY$, the coordinates of P'_i in the (X', Y', Z') coordinate system are computed as follows:

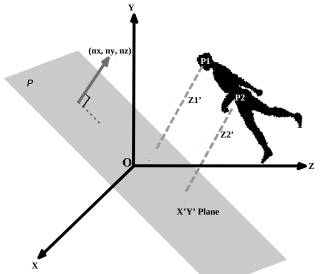


Fig. 2. Coordinate system transformation for silhouette alignment

$$Z'_i = \frac{AX_i + BY_i - Z_i}{\sqrt{A^2 + B^2 + 1}}, \quad X'_i = X_i - (Z'_i)(nx), \quad Y'_i = Y_i - (Z'_i)(ny). \quad (1)$$

Thus, the $X'Y'$ plane in the new $(X'Y'Z')$ coordinate system is actually the plane $\mathcal{P}$.

Volumetric Reconstruction from Point Cloud The proposed key pose estimation and gait recognition procedures, as described in Sections 3.2 and 3.3, respectively, require the point cloud to be mapped to a three dimensional voxel volume *V*. This makes it convenient to extract the relevant features by raster scanning the three dimensional volume along the width, height and depth dimensions. The volume is constructed by mapping the (X', Y', Z') coordinates of each object point present in the aligned point cloud into appropriate voxel positions within the volume. Suppose, *M*, *N*, *P*, respectively represent the dimensions of the voxel volume V along its width, height and depth directions. Also, let $\mathcal{M}_x$ and m_x respectively denote the maximum and minimum X' coordinates of the object points present in the aligned point cloud. Similar notations are used to denote the magnitudes of the maximum and minimum object point coordinates corresponding to the Y' and Z' directions of the transformed coordinate system. The mapped coordinates (X'_{i_v}, ${Y_{i_v}}'$, ${Z_{i_v}}'$) within the volume corresponding to the i^{th} object point (X'_i, ${Y_i}'$, ${Z_i}'$) is then computed as:

$$X_{i_v}{}'{=}\frac{X'_i - m_x}{\mathcal{M}_x - m_x}(M-1),\, Y_{i_v}{}'{=}\frac{Y'_i - m_y}{\mathcal{M}_y - m_y}(N-1),\, Z_{i_v}{}'{=}\frac{Z'_i - m_z}{\mathcal{M}_z - m_z}(P-1). \tag{2}$$

Fig. 3(a) shows the plot of a three dimensional voxel volume constructed from an aligned point cloud.

3.2 Extraction of Key Poses and Mapping of a Gait Sequence into Key Poses

A sequence of key poses represents a human gait cycle [2, 6, 11] (refer to Fig. 5). A sufficiently large number of walking sequences would help in accurate estimation of these key poses. Here, we describe a procedure for deriving a fixed number (K) of key poses from the gait sequences of a large number of subjects. The voxel volume V consists of either object voxels or non-object voxels. Let us suppose that each of the object voxels has been assigned a value of '*1*', whereas, each non-object voxel is assigned a value of '*0*'. Since, the depth information provided by Kinect is inherently noisy, V also contains a significant amount of noise, as seen in Fig. 3(a). Deriving meaningful features for key pose extraction requires an effective mechanism for smoothing the noisy volume.

Distance transform [15] labels each voxel within a binary volume with the Euclidean distance to the nearest object pixel. In the present context, for smoothing the volume *V*, we use a variant of the distance transform operator. Suppose d_{max} is the maximum value within the distance transformed volume. Then the value assigned to a non-object voxel having a distance transformed value of d is (1-$\frac{d}{d_{max}}$).

On application of the above operation, the value assigned to each non-object voxel in V lies within the range (*0*, *1*), so that a value closer to '*1*' indicates a smaller magnitude of d. On the other hand, the same value (i.e., '*1*') is retained corresponding to each of the object voxels. In addition to preserving the shape information, this step can effectively fill up all the noisy regions/ holes within the 3D aligned silhouette, thereby smoothing the volume. This helps in the extraction of robust features for key pose estimation as well as gait recognition, even if incorrect alignment occurs after the application of the alignment operation described in Section 3.1 due to noisy data. However, if the magnitude of d is high enough, then the voxel value in V is not altered. Fig. 3(b) shows the re-assigned values within the voxel volume V after applying the above noise removal procedure on Fig. 3(a).

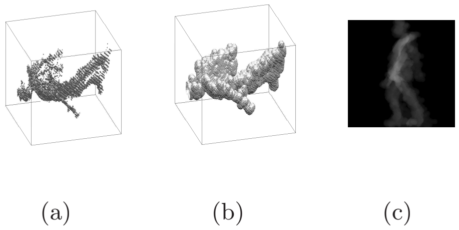


Fig. 3. (a) Aligned point cloud of a silhouette (b) Point cloud after noise removal and voxel filling (c) Average silhouette of the noise-free point cloud on the $Y'Z'$ plane

Since fronto-parallel view of gait contains the most informative gait features [7], we construct an equivalent fronto-parallel view silhouette frame using the depth information of the frontal surface of each silhouette captured by the Kinect. It is apparent from the above discussion that the $Y'Z'$ plane of the volume V provides information about the fronto-parallel view of a silhouette. In the present context, we propose to extract the feature vector for deriving key poses using a set of two dimensional silhouettes on the $Y'Z'$ plane. Each pixel within a two dimensional silhouette on the $Y'Z'$ plane is assigned a value equal to the mean of the values of all the voxel points whose projection on the $Y'Z'$ plane gets mapped to this pixel. Thus, if $(X'_{i_v}, Y'_{i_v}, Z'_{i_v})$ denotes the coordinates of a voxel point in the volume V and if $\mathcal{I}_{Y'Z'}$ denotes the projected frame on the $Y'Z'$ plane, then, the value assigned to the pixel (Y'_j, Z'_j) in the projected frame is given by:

$$\mathcal{I}_{Y'Z'}(Y'_j, Z'_j) = \frac{1}{M}\sum_{k=1}^{M} V(X'_{k_v}, Y'_{j_k}, Z'_{j_k}). \tag{3}$$

The silhouette thus obtained on the $Y'Z'$ plane after application of the above averaging operation is termed as the average silhouette. The average silhouette on the $Y'Z'$ plane derived from the point cloud of Fig. 3(b) is shown in Fig. 3(c).

The cluster centers obtained after clustering the silhouette sequence $\mathcal{I}_{Y'Z'}$ on the $Y'Z'$ plane are termed as the key poses in a gait cycle. In contrast to [2], where only binary silhouette sequences were used for key pose generation, the proposed method effectively captures the shape information of the fronto-parallel view as well as some dimensional information of the body parts corresponding to the frontal view. The value of K used in K-Means clustering is next determined from a rate distortion plot shown in Fig. 4. The plot shows the average distortion of the clustering operation as a function of the number of key poses. The distortion coefficient is plotted as the sum of the Euclidean distances between each silhouette sequence vector $\mathcal{I}_{Y'Z'}$ from its nearest cluster center. It is seen from the figure that the curve attains a minimum value for $K = 13$ and remains stable after $K = 14$. Hence, selecting the value of K as 13 seems to be a good choice for the estimation of key poses. Fig. 5 shows the thirteen representative key poses in a gait cycle obtained after the application of constrained K-Means clustering on the fronto-parallel view silhouettes on the $Y'Z'$ plane. Given an input silhouette sequence, a local sequence alignment procedure based on dynamic programming [16] is used to find correspondences between the frames of the sequence and the set of derived key poses. As an initial step, the alignment operation requires determination of a similarity score value [2] between a frame and each of the derived key poses, which is accomplished by computing the Euclidean distance between these two. The state transition information used in the alignment procedure can be stated as follows: if a certain frame of a sequence corresponds to a key pose k, then its succeeding frame must be mapped to either of key pose k or key pose $((k{+}1)$ modulo $K)$, $k = 1, 2, 3, ..., K$.

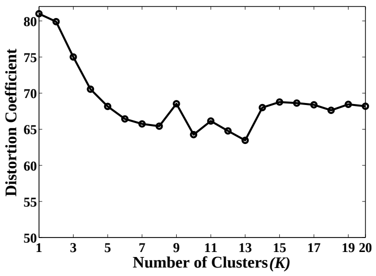


Fig. 4. Rate distortion curve for determining the appropriate number of key poses

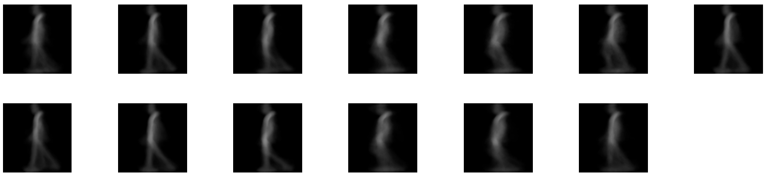


Fig. 5. Thirteen key poses derived for representing a gait cycle

3.3 Extraction of Gait Feature

Similar to the feature vector construction procedure for determining the key poses (refer to Section 3.2), extraction of the gait features is again done by considering the average silhouettes on the $Y'Z'$ plane. It may be noted that all the pixels belonging to a silhouette on this plane do not convey significant information about the gait of a subject. In most of the existing gait recognition literature [1, 2, 6], principal component analysis is used to reduce the feature vector length by eliminating redundant feature attributes. But this requires computation of the eigen silhouettes [2] corresponding to each frame of the sequence, which is time intensive.

It appears that pre-determination of the set of pixel coordinates carrying useful gait information can expedite the recognition procedure. This is accomplished by making use a variance image constructed from the aligned silhouette sequences corresponding to a large number of subjects on the $Y'Z'$ plane. The variance image is derived by computing pixel-wise variances of these sets of aligned silhouettes and is shown in Fig. 6(a). It is to be noted that the variance image actually preserves useful information about those pixel locations within an aligned silhouette frame which undergo significant change during walking. As seen in this figure, many pixels within the variance image have negligible variance, indicating that the silhouette points corresponding to those locations do not undergo significant position variation with respect to the silhouette center during walking. It is evident that these pixels carry little gait information and leaving out these pixels during the gait feature vector construction does not adversely affect the accuracy of recognition. In order to select only a specific set of pixels with important gait information, we binarize the variance image using an experimentally determined threshold β, so that only pixels with variances greater than β are retained in the binarized image. Figs. 6(b), (c), (d), (e) and (f) show the binary images obtained after thresholding the variance image, where the β values are set to 0, 0.1, 0.2, 0.3 and 0.4, respectively. From the figure, it is seen that the silhouette shape information is preserved at a high resolution corresponding to β values of 0.1 and 0.2, and hence, features derived considering these values of β are expected to contain significant gait information. A β

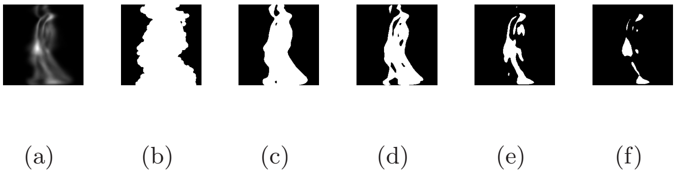


Fig. 6. (a) Variance image computed from the silhouette sequences of a large number of subjects (b), (c), (d), (e), and (f) Variance image binarization with $\beta = 0$, 0.1, 0.2, 0.3, and 0.4, respectively

value of 0 provides useful gait information along with redundant information to a certain degree. However, β values greater than 0.2 misses pixels (or elements) with significant gait information and should not be considered for gait analysis. We denote this β thresholded binary image by $\mathcal{I}_{\beta}$.

Given an input silhouette sequence, and the mapping of each of its frames to the appropriate key poses, we next extract a gait feature vector corresponding to each of the K key poses. The final gait feature is the concatenation of the feature vectors derived for each key pose. Consider a total of $\mathcal{N}$ subjects in the dataset and suppose l frames, starting with frame index t up to frame index $t+l-1$, namely, $F^k_{n,t}$, $F^k_{n,t+1}$, $F^k_{n,t+2}$, ..., $F^k_{n,t+l-1}$, of a gait sequence of the n^{th} subject are mapped to a key pose k, where n = 1, 2, 3, ..., $\mathcal{N}$, and k = 1, 2, 3, ..., K. Let $\mathcal{F}^k_{n,t+j}$ denote the vector of the most informative pixel values extracted from the frame $F^k_{n,j}$, for each j = 0, 1, 2, ..., l-1. It may be noted that the vector $\mathcal{F}^k_{n,t+j}$ is constructed using information only from those pixel locations that undergo significant variation during walking, as depicted in $\mathcal{I}_{\beta}$. Thus, a lower value of β will cause $\mathcal{F}^k_{n,t+j}$ to have a higher dimension, and vice-versa.

Then, the gait feature vector $\mathcal{G}^k_n$ corresponding to the k^{th} pose of the n^{th} subject is derived as follows:

$$\mathcal{G}^k_n = \frac{1}{l}\sum_{j=0}^{l-1}\mathcal{F}^k_{n,t+j}, \quad n = 1, 2, 3, ..., \mathcal{N}, \quad k = 1, 2, 3, ..., K. \tag{4}$$

3.4 Recognition of a Test Subject using the Proposed Feature

We denote the $\mathcal{N}$ subjects in the training set as S_1, S_2, S_3, ..., and $S_{\mathcal{N}}$. Let $\mathcal{G}^k_{n,tr}$ denote the feature vector corresponding to the k^{th} pose of the n^{th} training subject, where n = 1, 2, 3, ..., $\mathcal{N}$ and k = 1, 2, 3, ..., K. A similar notation $\mathcal{G}^k_{te}$ is used to denote the feature vector corresponding to the k^{th} pose of an input test subject. For each of the $\mathcal{N}$ subjects in the training set, a measure of similarity is next computed which signifies the likelihood of the test subject to belong to the class of the current training subject. Suppose, out of the total number of K key poses, only p of them k_1, k_2, k_3, ..., k_p, are common for a given combination of training and test sequences. Initially, the feature vectors derived corresponding to this set of matching p key poses are concatenated to form a single vector. Thus, if $\mathcal{G}_{n,tr}$ and $\mathcal{G}_{te}$, respectively denote these concatenated feature vectors corresponding to the n^{th} training subject and the given test subject, then:

$$\mathcal{G}_{n,tr} = \begin{bmatrix} \mathcal{G}^{k_1}_{n,tr} \\ \mathcal{G}^{k_2}_{n,tr} \\ \mathcal{G}^{k_3}_{n,tr} \\ \cdots \\ \mathcal{G}^{k_p}_{n,tr} \end{bmatrix} \qquad and \qquad \mathcal{G}_{te} = \begin{bmatrix} \mathcal{G}^{k_1}_{te} \\ \mathcal{G}^{k_2}_{te} \\ \mathcal{G}^{k_3}_{te} \\ \cdots \\ \mathcal{G}^{k_p}_{te} \end{bmatrix}.$$

It is to be noted that each of $\mathcal{G}_{n,tr}$ and $\mathcal{G}_{te}$ might consist of missing attribute (null) values corresponding to the key poses those are absent in the training and the test sequences, respectively. Also reconstruction of the feature vectors by estimating these null attribute values is difficult because of the availability of insufficient number of sequences. Thus, the measure of similarity must be computed by comparing only the non-null attribute values common to both the vectors $\mathcal{G}_{n,tr}$ and $\mathcal{G}_{te}$. Since different pairs of training and test sequences will have different sets of non-null matching attributes, it is necessary to normalize the similarity metric to make it independent of the magnitudes of the individual attributes. The 'cosine' similarity metric is beneficial in such cases. The cosine similarity $\mathcal{D}_n$ between the vectors $\mathcal{G}_{n,tr}$ and $\mathcal{G}_{te}$ is computed as:

$$\mathcal{D}_n = 1 - cos(\alpha), \tag{5}$$

where, α is the angle included between the two vectors, given by:

$$\alpha = \frac{\mathcal{G}^T_{n,tr}\mathcal{G}_{te}}{||\mathcal{G}_{n,tr}||||\mathcal{G}_{te}||}. \tag{6}$$

The test subject is identified as S_r if:

$$\mathcal{D}_r \leq \mathcal{D}_n, \forall n = 1, 2, 3, ..., \mathcal{N}. \tag{7}$$

4 Experimental Evaluation

There is no existing database that provides the depth information of the gait of subjects using depth cameras like Kinect. Hence, to test the effectiveness of our approach, we construct a new dataset[1]. In the following sub-sections, we provide an elaborate description of the experimental setup, the testing protocol and, finally, an extensive evaluation of the proposed method using the captured dataset.

4.1 Dataset Description

A total of 29 subjects have been used in building our database. The experimental setup for recording both the training and the test sequences is made similar to the one shown in Fig. 1. A Kinect camera (K) in combination with the SDK provided by Microsoft is used for collecting the datasets. The camera is positioned at a height of 2.5 metres from the ground over a narrow pathway, facing downwards. The tilt angle of K is set to -23°. As a subject passes through this zone, the real-world X, Y and Z coordinates of the points on the frontal surface of the silhouette of each subject as tracked by K are recorded.

For each subject, we record two distinct sequences *T1* and *T2*, which are used as training sets in our experiments. Test sets *T3* and *T4* for each subject are respectively collected under two different frame rates: 30 fps and 15 fps. Thus, in total, we have 116 distinct frontal sequences, containing four sequences corresponding to each subject. Due to a limitation on the maximum depth sensing range of Kinect, which is only 4 metres, many of these recorded sequences lack complete gait cycle information. Table 1 presents a statistics of the recorded training and test sets, showing the percentage of sequences that have missed k out of the K (= 13) key poses, k = 1, 2, 3, ..., 13.

[1] Available on request

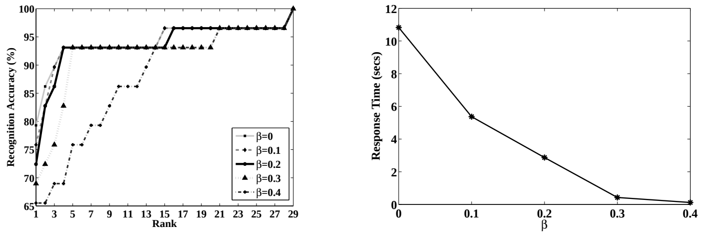
Table 1. Percentage of key poses missed by various sequences

Dataset	Key Pose Indices													
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
T1	17.24	24.14	06.89	03.45	06.89	03.45	13.79	20.70	03.45	00.00	00.00	00.00	00.00	00.00
T2	17.24	00.00	06.89	20.70	06.89	00.00	06.89	20.70	00.00	17.24	00.00	00.00	03.45	00.00
T3	10.34	17.24	06.89	06.89	06.89	03.45	20.70	27.60	00.00	00.00	00.00	00.00	00.00	00.00
T4	10.34	13.79	10.34	06.89	06.89	03.45	20.70	27.60	00.00	00.00	00.00	00.00	00.00	00.00

4.2 Testing Protocol and Results

Experiments are conducted in the context of biometric based identification where the gait feature derived from a test sequence is compared against a gallery of features derived from a number of training subjects. Implementation of the proposed algorithm is done in MatLab environment (version R2011a) on a system having 2.50 GHz Intel Core i5 processor and 4GB RAM.

First, we experimentally determine an optimal value of β required to binarize the variance image (refer to Fig. 6). For this, we plot cumulative match characteristic (CMC) curves corresponding to β values of 0, 0.1, 0.2, 0.3, 0.4 in Fig. 7(a), using only *T1* as the training set. The recorded response times for these different β values are plotted in Fig. 7(b). From Fig. 7(a), it is seen that the proposed method has a high recognition rate for $\beta \leq 0.2$, even in the presence of incomplete cycle sequences. But as observed from Fig. 7(b), the processing times required for $\beta = 0$ and $\beta = 0.1$ are significantly high. On the other hand, the response time corresponding to $\beta = 0.2$ is at most 3 seconds which is reasonably fast. Hence, the choice of the value of β as 0.2 can be considered as an effective balance between processing time and recognition accuracy. Each of the subsequent experiments conducted considers $\beta = 0.2$.



(a) (b)

Fig. 7. (a) Cumulative match characteristic curves showing variation of recognition accuracy with rank for different values of β (b) Response times of the algorithm corresponding to these β values

It is expected that an increased volume of training data will help in achieving higher accuracy during test cases. This is experimentally verified in Table 2. The table shows the recognition performance corresponding to the test sets *T3* and

T4, recorded at 30 fps and 15 fps, respectively, and in presence of only *T1*, only *T2*, and both *T1* and *T2* (*T1*+*T2*) as training sets. Percentage accuracies in the table are shown using two different similarity measures: Cosine (*D1*) and Euclidean (*D2*).

Table 2. Variation of recognition accuracy with frame rate for different training data set combinations and for Cosine (*D1*) and Euclidean (*D2*) similarity measures

	T3		*T4*	
Training Set	*D1*	*D2*	*D1*	*D2*
T1	72.41	31.03	72.41	27.59
T2	55.17	13.79	51.72	13.79
T1+*T2*	86.21	41.38	79.31	37.93

The advantage of using a normalized similarity measure, such as the Cosine metric, in comparing feature vectors with missing attributes is evident from the table. Moreover, from Tables 1 and 2, it is seen that there is no significant variation in recognition performance with reduction in frame rate, as long as the available key poses corresponding to the two frame rates closely match each other. In general, recognition rate is not remarkably high when only *T2* is chosen as the training set. This is primarily because the gait sequences present in *T2* contain a higher percentage of missing key poses (refer to Table 1).

To evaluate the effectiveness of an algorithm, it is often required to determine if its performance is satisfactory for a sufficiently small value of rank. A test subject is said to be perfectly classified at a given rank *r*, if the correct class of this subject is one of the top *r* predictions of the algorithm. We plot a rank-wise improvement in classification performance of the proposed method in Fig. 8 corresponding to each of the training sets *T1*, *T2* and *T1*+*T2*. It is seen from the figure that using both *T1* and *T2* as training sets, a recognition rate greater than 90% is achieved within a rank of 4, which highlights the efficacy of the proposed method in gait recognition setups similar to that shown in Fig. 1.

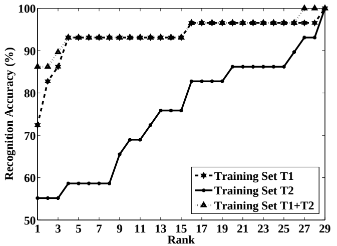


Fig. 8. CMC curves showing improvement in recognition rate with rank for the different training sets

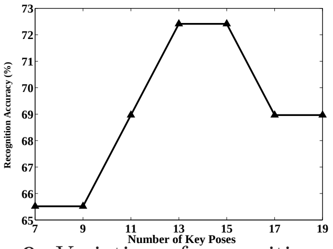


Fig. 9. Variation of recognition accuracy with number of key poses using *T1* as the training set

In each of the previous experiments, the value of K has been set to 13 (determined from the rate distortion plot of Fig. 4). However, a reader might be interested in studying the effect of change of the number of key poses (K) on the recognition rate. Hence, we plot the variation in percentage accuracy corresponding to K values of 7, 9, 11, 13, 15, 17, and 19 in Fig. 9. Training set for this experiment consists of only *T1*. It is seen from the figure that the curve initially has a non-decreasing trend for values of $K \geq 7$. It attains a peak value at $K = 13$ and again decreases for values of $K > 15$. Thus, the choice of the value of K as 13 in each of the previous experiments is rightly justified. The reduction in recognition rate corresponding to $K \geq 7$ is due to the higher number of missing key poses in the gait sequences.

Finally a comparative performance analysis of the proposed method is made with existing work on frontal gait recognition using Kinect, namely, GEV [5], PDV [6], skeleton-covariance feature [9] and a pose based skeleton feature [11]. The effectiveness of the use of depth data in gait recognition is studied by comparing our approach with some of the traditional gait recognition methods which use RGB cameras for data collection, namely, GEI [1] and PEI [2]. Results are shown in Table 3 using only *T1* as the training set. It can be seen that the proposed approach outperforms each of the state-of-the-art gait recognition techniques by more than 20%, which is remarkable. The slightly higher response time of our algorithm as compared to [1, 2, 5, 9, 11] can be sacrificed for achieving the significant improved recognition rate. This superior recognition performance together with a fast response time indicate the potentiality of this method in performing gait recognition from incomplete cycle sequences.

Table 3. Comparative performance of the proposed method with the existing literature

Gait Recognition Algorithms	Accuracy (%)	Time (in secs)
GEV [5]	27.59	2.58
PDV [6]	51.72	15.62
Skeleton Co-Variance Feature [9]	34.48	1.08
Skeleton Pose Based Feature [11]	51.72	2.23
GEI [1]	31.03	0.42
PEI [2]	44.83	1.34
Proposed feature	72.41	2.87

5 Conclusion and Future Scope

From the extensive set of experiments described in Section 4.2, it can be concluded that the proposed gait recognition procedure can be potentially applied in surveillance sites similar to Fig. 1. The pose based approach helps in preserving kinematic details in recognizing the gait of a subject from a given sequence. Evaluating the performance of the proposed algorithm in presence of a larger number of subjects and combining both back and front view sequences in the recognition procedure would be a direction for future research.

References

1. J. Han and B. Bhanu. Individual Recognition Using Gait Energy Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006.
2. A. Roy, S. Sural, and J. Mukherjee. Gait Recognition Using Pose Kinematics and Pose Energy Image. *Signal Processing*, 92(3):780–792, 2012.
3. E. Zhang, Y. Zhao, and W. Xiong. Active Energy Image Plus 2DLPP for Gait Recognition. *Signal Processing*, 90(7):2295–2302, 2010.
4. C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian. Frame Difference Energy Image for Gait Recognition with Incomplete Silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.
5. S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes. Gait Energy Volumes and Frontal Gait Recognition Using Depth Images. In *International Joint Conference on Biometrics*, pages 1–6, 2011.
6. 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*, 25(1):53–63, 2014.
7. N. V. Boulgouris, D. Hatzinakos, and K. N. Plataniotis. Gait recognition: A Challenging Signal Processing Technology for Biometrics Identification. *IEEE Signal Processing Magazine*, 22(6):78–90, 2005.
8. Z. Zhang. Microsoft Kinect Sensor and Its Effect. *IEEE Multimedia*, 19(2):4–10, 2012.
9. M. S. N. 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*, article number 20. ACM, December 2012.
10. M. Milovanovic, M. Minovic, and D. Starcevic. Walking in Colors: Human Gait Recognition Using Kinect and CBIR. In *IEEE Multimedia*, 20(4):28–36, 2013.
11. P. Chattopadhyay, S. Sural, and J. Mukherjee. Gait Recognition from Front and Back View Sequences Captured Using Kinect. In 5^{th} *International Conference on Pattern Recognition and Machine Intelligence*, pages 196–203, 2013.
12. J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-Time Human Pose Recognition in Parts from Single Depth Images. *Communications of the ACM*, 56(1):116–124, 2013.
13. M. Hofmann, S. Bachmann, and G. Rigoll. 2.5D Gait Biometrics Using the Depth Gradient Histogram Energy Image. In 5^{th} *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 399–403, 2012.
14. D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2012.
15. J. Wang, Y. Makihara, and Y. Yagi. Human Tracking and Segmentation Supported by Silhouette-Based Gait Recognition. In *IEEE International Conference on Robotics and Automation*, pages 1698–1703, 2008.
16. L. R. Rabiner. A Tutorial on Hidden Markov models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.