Multimedia Tools and Applications (2021) 80:10141–10166
https://doi.org/10.1007/s11042-020-10132-z

Robust gait based human identification on incomplete and multi-view sequences

Utkarsh Shreemali[1] · Anirban Chakraborty[1]

Received: 5 May 2020 / Revised: 23 September 2020 / Accepted: 23 October 2020 /
Published online: 17 November 2020

Abstract

Gait based person identification is an important research area in the field of video surveillance. The major challenges faced by gait recognition systems in real-life scenarios include view variance, occlusion and resultant unavailability of a complete sequence containing a gait cycle. In this work, we propose a novel robust gait recognition framework capable of handling these challenges. We show how Gait-Energy-Images (GEIs) can be accurately constructed from largely incomplete input silhouette sequences. This provides an immediate advantage over current literature that assumes availability of complete sequences. We then highlight the shortcoming of most of the current view-invariant models that perform sub-optimal transformation of probe and gallery sequences captured in different views for comparison. We propose a model which jointly estimates and learns the optimal transformation for comparison of probe and gallery GEIs. Through extensive experiments, we show that our proposed framework is able to outperform most state-of-the-art methods on multiple benchmarks.

Keywords Video surveillance · Gait based person recognition · Person re-identification · View invariance · Missing data

1 Introduction

The task of identifying humans within and across cameras has been a major topic of interest in the video surveillance community. In a general setting of a network of cameras, given a sequence of observations of a target person, the aim is to identify or re-identify that person

Utkarsh Shreemali was a graduate student at the Dept. of Computational and Data Sciences, Indian Institute of Science, Bangalore when this research work was carried out as a part of his M.Tech. dissertation.

✉ Anirban Chakraborty
anirban@iisc.ac.in

Utkarsh Shreemali
utkarshs@iisc.ac.in

[1] Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

across cameras. This task falls under the domain of sequence-to-sequence matching. The sequences observed across cameras tend to have huge variations due to change in viewpoints, presence of occlusions, illumination and pose variations, etc.

Gait is defined as a person's manner of walking. It is an important biometric property since it is difficult to fake and is contact-less, which gives it advantage over other biometric based human recognition methods like iris and finger-print recognition. Thus, gait analysis has great significance in the field of video surveillance for the task of person identification. When a person walks, he/she repeats the walking actions periodically. One such period capturing all the walking actions of a person is called a gait cycle. Gait analysis models work on a sequence of observations containing the entire gait cycle of a person. A single gait cycle is considered a sufficiently good representation of the entire sequence containing the gait cycle since the remaining portion in the sequence is generally just a repetition of the information already contained in a gait cycle.

One main aspect of gait analysis is that it does not make use of chromatic appearance of a person for identification. Most models of gait analysis make use of either the pose information [46] extracted from the video sequence or take silhouettes [51, 53] as inputs. In the latter case, gait analysis models make use of features like Gait Energy Image (GEI) [15], Gait Entropy Image (GEnI) [4], Gait Flow Image (GFI) [29], Period Energy Image (PEI) [17], etc. Out of these Gait Energy Images (GEIs) are found to be most effective in performing robust gait recognition [23]. GEIs are obtained by averaging all the silhouettes in an observed gait cycle of a person.

In practical scenarios, gait based systems need to handle the presence of occlusions and noise (due to the presence of another person or object in the frame) in the captured video sequence of a person. In such cases, the noisy/occluded frames are rendered unusable, and hence a clean silhouette sequence containing the complete gait cycle of an individual may not be available. Despite this being a crucial challenge in gait based identification, only a handful of works have looked into it. Another major challenge comes from the cross-camera scenario where such a system is deployed. In many scenarios, cameras in the network have wide variation in viewpoints. Although this problem is well-known and present in all cross-camera identification problems, its impact is felt worst in the gait-based identification problems-mainly because of the absence of chromatic appearance information.

The objective of this work, therefore, is to build a robust gait based person recognition system that works with a temporal sequence of observations/silhouettes from a given target in each camera and is robust to the challenges associated with such problems, viz., missing data and view variability. Figure 1 gives a brief overview of our method. The major contributions of this work are as follows:

1. First, we look at the missing data problem and propose a model which takes the incomplete sequence of silhouettes and constructs a GEI which can be used for the task of gait based person identification. The aim of this module is to construct a GEI out of a smaller subset of silhouettes, which would appear visually identical to the GEI constructed from complete silhouette sequence.
2. In the second part of our work we aim towards making our gait-based identification framework view-invariant. Given a pair of probe and gallery GEI in different view angles, we propose a novel method which estimates an optimal view most suitable for comparing these GEIs. For this we also learn a view transformation model so that the probe and gallery GEIs can be transformed to the optimal view.

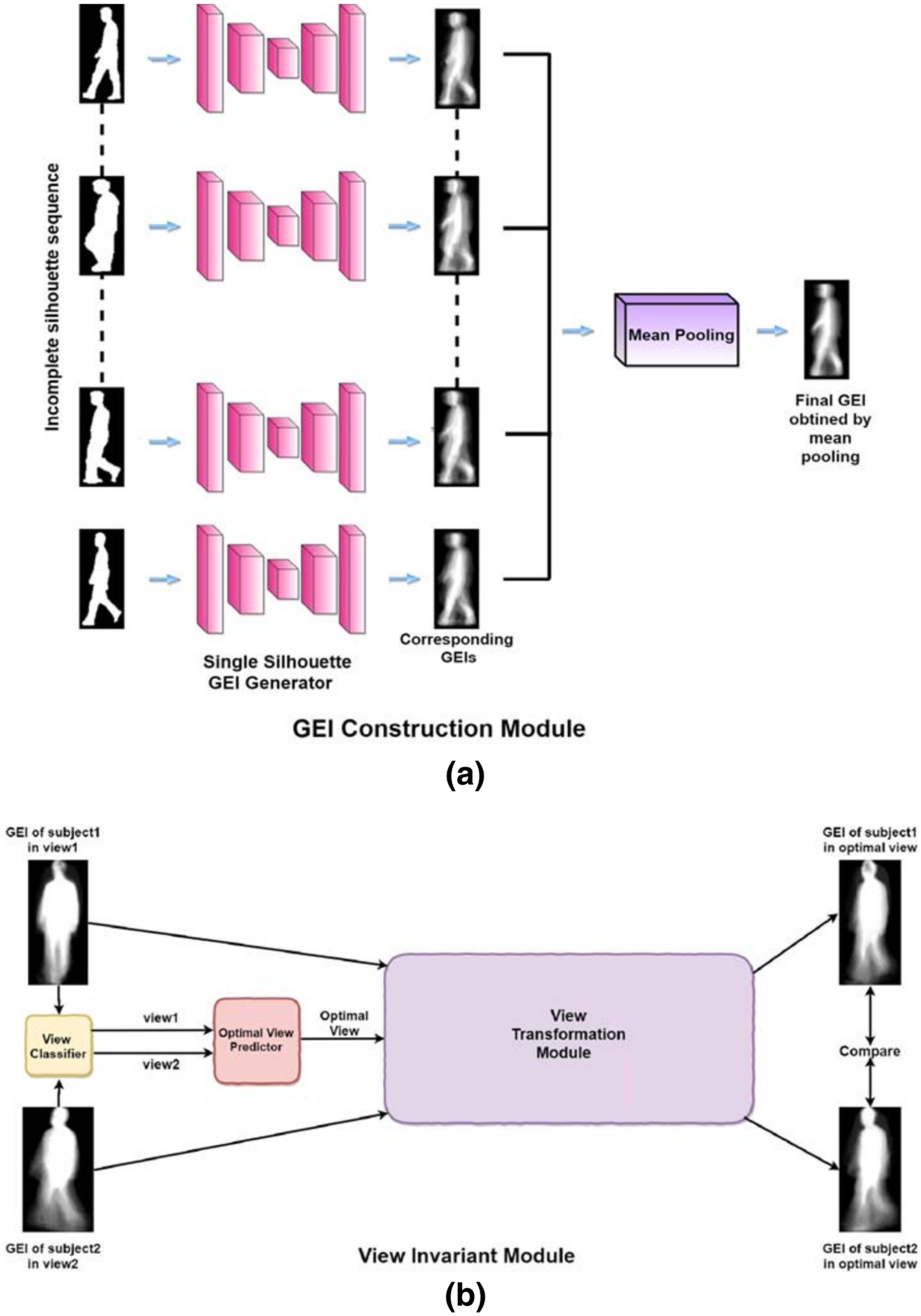

Fig. 1 Overview of our method: **a** The GEI Construction Module takes the incomplete sequence of silhouettes as input and generates a GEI corresponding to each silhouette which are then mean pooled to give a final GEI for the input sequence. **b** The View Invariant Module uses a view classifier to predict the view angle of probe and gallery GEI at test time. Based on view classifier's outputs, the optimal view predictor predicts the optimal view, to which the probe and gallery GEIs are transformed for better comparison. The optimal view and view transformation are jointly estimated in an iterative fashion

3. We perform extensive experiments on two popular benchmark datasets, viz., CASIA-B [55] and OU-ISIR [23] and show that the proposed framework can outperform most state-of-the-art methods on the task of gait-based person identification.

The rest of the paper is organized as follows. In Section 2 we will review the related work in the field of gait based person recognition. In Section 3, we give our proposed approach to handle the problem of missing data and view variance. Section 4 contains the experimental details whereas the results are presented in Section 5. Finally, we conclude our work in Section 6.

2 Related work

Gait, being an important bio-metric property of a human, is used in different branches of research [11]. Two important areas are - Gait Analysis and Gait based recognition. Gait analysis is a field where human motion is studied and important parameters that characterize human gait events are determined. One such area is bio-mechanical analysis of gait [42] that is used in health diagnostics and performance analysis in sports. This includes modelling the human gait cycle [39, 44]. In [41], Patil et al. use Extreme Machine Learning (ELM) [22] for clinical gait analysis. Gait of a person is periodic and can be divided into 8 sub phases [41]. A gait cycle last for about 1–1.12 seconds [43]. Several of the gait analysis systems use sensors like Inertial Measurement Units (IMUs) including accelerometers, gyrosensors, force sensors, strain gauges, inclinometers, goniometers, etc. Apart from these, vision based gait analysis systems have also shown to give good results. González et al. [13] and Kyrarini et al. [28] provide a comparative study of vision based gait analysis systems and wearable sensors based systems. While vision based systems give acceptable results but the error rate is slightly higher than IMU based methods. However, vision based systems have the advantage of being relatively cheaper, easier to monitor and contact-less. Most of the sensors based analysis requires good quality data in sufficient quantity captured using different types of sensors.

In gait based recognition, the human gait is studied to find distinctive features of a person to uniquely identify him/her from others. Our work falls in the domain of gait based recognition. Most of the gait recognition methods can be classified into two categories: model based methods [6–9] and appearance based methods [12, 26, 27, 31]. Model based methods model the underlying structure of the human body whereas appearance based methods try to extract gait features from gait cycles. Our work falls under the latter category. In many situations, it is difficult to precisely model the human body structure from videos captured under challenging real-life conditions. This gives appearance based methods an advantage over model based methods as they are capable of performing on such challenging videos/observations. The commonly used features in appearance based methods are Gait Energy Image (GEI) [15], Gait Entropy Image (GEnI) [4], Gait Flow Image (GFI) [29], Chrono-Gait Image (CGI) [50], etc. Gait Energy Image (GEI) is formed by averaging the size-normalized and centre aligned silhouettes of a gait cycle of person. Gait Entropy Image (GEnI) is formed by calculating Shannon entropy for each pixel in the silhouettes of a sequence. Gait Flow Images (GFIs) are formed by determining the optical flow field from silhouettes of a gait cycle. Chrono-Gait Images (CGIs) encode the temporal information among the silhouettes of a gait cycle using a colour mapping. GEIs are found to be most effective for the task of person identification based on gait [23].

Most of the works done in the field of gait recognition assume the availability of silhouette sequence containing complete gait cycle. However, in real world scenarios it is very rare to get a complete sequence and hence this becomes a major limitation of the current literature in this field. To the best of our knowledge, the only work that tries to address this problem is by Babaee et al. [2]. They propose a fully convolutional network, named ITCNet(Incomplete to Complete GEI network) which transforms a GEI obtained from incomplete silhouette sequence(incomplete GEI) to a complete GEI. ITCNet performs this transformation in a progressive manner - an initial incomplete GEI is transformed to a partially complete GEI which is finally transformed to complete GEI. In [40], Ortells et al. propose a robust approach for gait recognition on noisy and occluded silhouettes. ITCNet and our model handles incomplete silhouette sequences obtained after removing noisy and occluded silhouettes from the sequence. We compare and contrast our model with ITCNet and discuss the reasons behind better performance of our model in Section 3.1.

Since viewpoint change is a major challenge in gait recognition, recent works focus on developing view invariant models. These works take one of the following three approaches: (1) creating a 3-D model of the human body [33] and obtain 2-D projections in different views [1, 34, 49, 56] (2) using view invariant features for gait recognition [12, 27], (3) learning cross view projectors [26, 36]. In Fig. 2, we provide a visual taxonomy of the approaches of gait based methods. In [35], Lua and Tjahjadi constructed a robust 3D representation of human body for abnormal gait behaviour recognition. The View Transformation Model (VTM) [5] is commonly used and is capable of transforming gait features from one view to another. Singular Value Decomposition (SVD) is used to compute the projection matrix and view-invariant features from a GEI. However, VTM can only transform from a specific angle to another and relies heavily on the performance of view angle estimator. In [25], Kusakunniran et al. formulate the problem of view invariant gait recognition as a regression problem and create a View Transformation Model using support vector regression. They overcome the problem of over-fitting in the original VTM by using truncated SVD and focus on local regions of interest as opposed to global features.

In [20], Hossain et al. propose to divide the human body into 8 sections with 4 overlapping ones and then extract features from sections which are more robust to clothing variations. In [45], Shiraga et al. proposed a CNN based gait recognition model, GEINet and produced the then state-of-the-art results. In [54], Yu et al. use stacked auto-encoders to

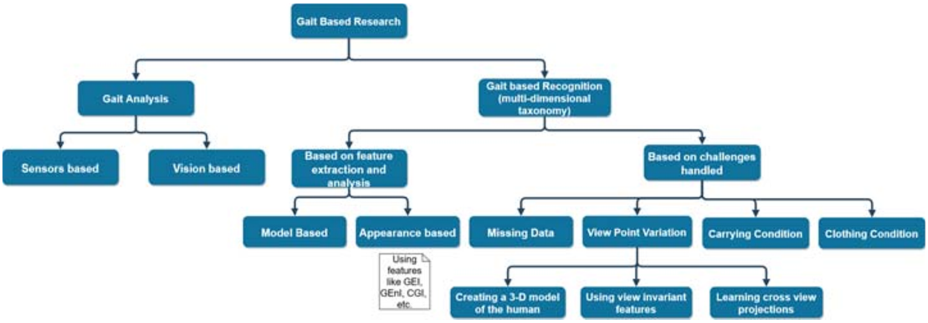
Gait Based Research
Gait Analysis
Gait based Recognition (multi-dimensional taxonomy)
Sensors based
Vision based
Based on feature extraction and analysis
Based on challenges handled
Model Based
Appearance based
Missing Data
View Point Variation
Carrying Condition
Clothing Condition
Using features like GEI, GEnI, CGI, etc.
Creating a 3-D model of the human
Using view invariant features
Learning cross view projections

Fig. 2 Visual taxonomy of the gait based approaches. Gait based research can be broadly classified into two areas - Gait Analysis and Gait based Recognition. While gait analysis involves studying and modelling the human motion, gait based recognition focuses on finding distinctive features to uniquely identify a person based on his/her gait

generate view invariant gait features in a progressive manner. They transform GEIs in any given view angle, and under any given clothing and carrying condition to a GEI in normal (90°) view and normal carrying and clothing conditions. The advantage of their method over VTM is that they use a single model to extract invariant gait features robust to view, clothing and carrying condition. In [37], Marín-Jiménez et al. showed that gait recognition under a multi-task learning set-up gave similar or superior performance to that under single-task learning set-ups. In [53], Yu, Shiqi, et al. propose a method named GaitGAN where they tackle the challenges of view variance, clothing changes and carrying condition variations using a single model. GaitGAN transforms GEI obtained from any given view angle(from 0° to 180°) and under any given clothing(i.e. with or without coat) and carrying condition (i.e. with or without bag) to a GEI with a view angle of 90° under normal clothing and carrying condition(i.e without a bag or coat). GaitGAN uses two discriminator networks for this purpose: one which ensures that the generator generates realistic GEI and the other which ensures that the generated GEI retains the human identification information.

In [52], Xing et al. propose a novel reformulation of Canonical Correlation Analysis(CCA), named Complete CCA which provides a stable and efficient alternative to CCA for solving the gait recognition problem. In [21], Hu et al. seek a unitary projection to project original gait features extracted from any view into a low dimensional feature space while improving the discriminative power for identification. Their method is named View-invariant Discriminative Projection (ViDP). The advantage of ViDP is that it can perform multi-view gait recognition without finding out the view in which the query sequence was captured. In [17], He et al. propose a multi-task GAN model which learns view-specific feature to perform gait recognition. Their model learns a view transformation layer which performs view transformation in a low-dimensional latent space. They also propose a new multi-channel gait template, called Period Energy Image (PEI). PEIs combine the benefits of using GEIs and CGIs into a single gait template. However, PEIs are not suitable to use when we do not have the complete sequence available since formation of good quality PEIs require sufficiently large number of silhouettes. This gives GEIs an upper hand in practical cases. In [10] Chao et al. propose to use silhouette sequence as a set rather than a sequence and achieve state-of-the-art performance. However, they propose a very heavy network with large number of parameters when compared to our much simpler network.

The current state-of-the-art methods handle the problem of view variance by transforming the probe sequence features to gallery view or to some pre-determined fixed view. However, such transformations are sub-optimal. In [38], Daigo et al. show that better accuracies can be achieved by considering an intermediate angle for transforming the probe and gallery gait features to that angle. However, in their work they do not propose any method to determine the intermediate angle. In this work, we propose a novel method to learn an optimal view for better comparison of probe and gallery observations.

3 Proposed approach

We present the proposed framework in two parts. First, we describe the model for GEI construction from incomplete silhouette sequences. Subsequently, we present our novel gait identification framework and show how the GEIs can be used for identification under large viewpoint variation.

3.1 Handling missing data

In this section, we propose our model to handle the problem of missing data. While capturing the video of a person, the silhouette sequence may often contain noisy frames because of partial/full occlusion. GEI constructed from such a sequence including these noisy silhouettes would lead towards a representation of poor quality. In such cases, the workaround is to discard the noisy silhouettes and estimate GEI from the remaining ones in the gait cycle. However, a GEI formed on such an incomplete subset often does not contain the information necessary for distinguishing individual's identity. To overcome this problem we propose a model which takes incomplete sequence of silhouettes as input and estimates corresponding GEI representative of all the silhouettes across the complete gait cycle (if it were available) and hence is suitable to perform the identification task. The number of available silhouettes in a sequence may vary and the worst case scenario is when we only have access to one silhouette in the complete gait cycle. Therefore, we propose a model, called **Single Silhouette GEI Generator**, that generates a GEI corresponding to each of the available silhouettes in the input sequence. These GEIs are then combined in ways described below to obtain the final GEI representative of the input sequence.

3.1.1 Network architecture

The architecture of Single Silhouette GEI Generator network is shown in Fig. 3. It takes each silhouette in a single input sequence as input and outputs a corresponding intermediate GEI. This intermediate GEI is a representation of the information contained in the entire gait cycle (if it was available). Finally, all these intermediate GEIs constructed from the silhouettes of the single incomplete sequence are average pooled along time to yield the final GEI corresponding to the entire silhouette sequence. We train the network using **mean square error loss** calculated between each pixel of the ground truth GEI and the GEI generated by our network as shown in (1) and (2). C represents the Single Silhouette GEI generator model, x_{sil} is one silhouette in the available sequence of length l (l may be less than a gait cycle) of a subject, $C(x_{sil})$ represents the GEI generated from x_{sil} and x_{GEI} is the ground truth GEI. $\mathbb{E}$ is the expectation. GEI_{final} is the final GEI and $x_{sil}^{(i)}$ represents the i^{th} silhouette in the sequence.

$$\min_{C} L_{MSE} = \mathbb{E}\|C(x_{sil}) - x_{GEI}\|_2 \tag{1}$$

$$GEI_{final} = \frac{1}{l}\sum_{i=1}^{l} C(x_{sil}^{(i)}) \tag{2}$$

An important aspect of our network is the use of skip connections [32]. The skip connections [32] allow the passage of appearance information from a shallow, fine layer to a deep, coarse layer where it is combined with the semantic information to produce better results. However, our model is structurally very different from [32]. Also, we use the network in Fig. 3 only to generate intermediate GEIs from silhouettes. The final GEI used for identification are only obtained after averaging all the intermediate GEIs. Another advantage of our model is that it is independent of the order in which the silhouettes appear in the sequence since we are averaging the intermediate GEIs generated from silhouettes to obtain the final GEI in the end. This allows our model to be independent of the temporal occurrence of the silhouettes in the sequence.

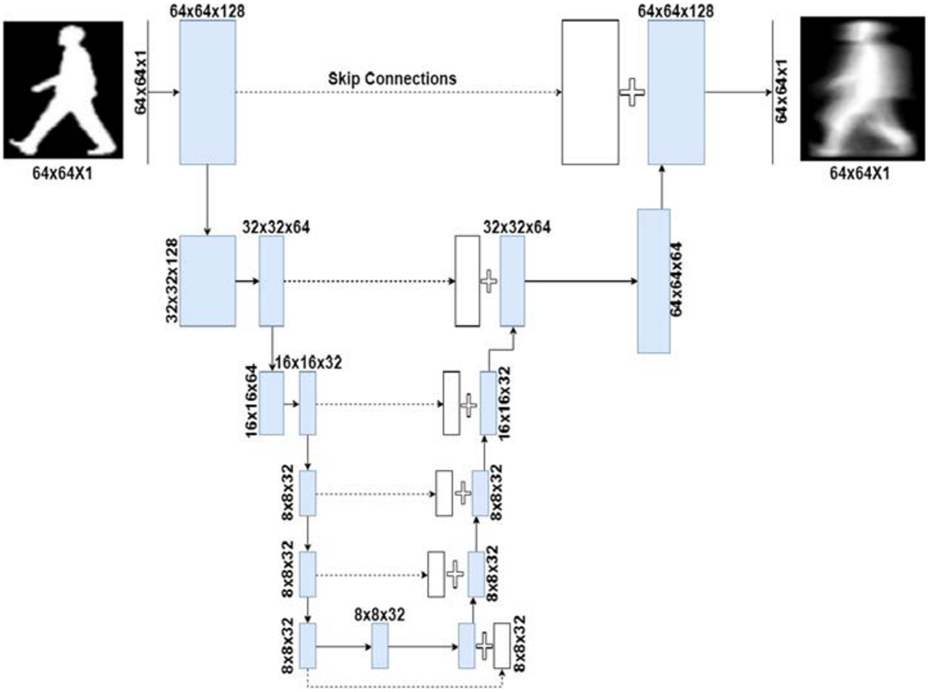

Fig. 3 Architecture of the Single Silhouette GEI Generator. The network takes a single size-normalized silhouette, of dimension 64×64, as input and generates an intermediate GEI corresponding to that particular silhouette. The same is done for each silhouette in the input sequence. Finally, all the generated intermediate GEIs are averaged together to give a single GEI representative of the entire input sequence of silhouettes (refer Fig. 1a). The dimensions over each block represents the dimensions of the corresponding output feature maps (for eg. HxWxC represents a feature map with height = H, width = W and number of channels = C). In the figure, the left half represents a series of convolution and pooling layers whereas the right half represents a series of de-convolution layers. In the bottom most layer of the figure, convolution operation is applied without changing the dimensions of the feature map using appropriate padding. Note that both silhouette and GEI are single channel images

The only other work looking into GEI reconstruction from few silhouettes is the ITCNet [2]. Hence, here we wanted to differentiate our model from that proposed in [2]. In fact, our network differs from ITCNet [2] in various aspects. ITCNet [2] builds a complete GEI from an incomplete GEI. The input to their network in an incomplete GEI formed by averaging a continuous silhouette sequence starting from a randomly selected silhouette from the complete sequence. Their network constructs complete GEI in a progressive manner. We construct a GEI in an end-to-end manner i.e. the input to our network is a single silhouette and the output is a GEI. After the GEIs from each of the silhouettes have been generated, we average out all these GEIs to get our final GEI. This allows us to preserve the most distinctive features of a person, which should be common across all generated GEIs, in the final GEI. The GEIs generated by a single silhouette may contain some noise/error. However, as these GEIs are averaged the contribution of error/noise reduces and the distinctive features of a person which are common across all the generated GEIs dominate the final representation i.e final GEI. Additionally, this approach allows us to work on discontinuous silhouette sequences since we are averaging in the end which does not depend on the order in which the GEIs are pooled. Another aspects that differentiates our network from ITCNet

[2] is that we use skip connections. This allows us to preserve the distinctive features of a person across the network and generate GEIs in an end-to-end manner. These aspects of our network enable it to perform better than ITCNet [2] (refer Section 5.1).

3.1.2 GEI pooling approaches beyond average pooling

Apart from the fore-mentioned method, we also tried some other approaches to obtain the final GEI from the intermediate GEIs, which gave good results when the number of available silhouettes is 10 or above, but performed poorly when the number of available silhouettes is in the range 1 to 10. The limitation of these methods is that we need to fix the length of the input silhouettes sequence and when the sequence is of smaller length we repeat a few silhouettes. However, the intuition and motivation behind these approaches is worth mentioning.

Weighted Pooling One of the approach was to do weighted pooling of the intermediate constructed GEIs, instead of simple averaging, to form the final GEI. The weights given to the intermediate GEIs while forming the final GEI can be learned by the network. The intuition behind this approach was that GEIs formed from silhouettes where legs and hands of a person are clearly visible should contribute more towards final GEI as compared to GEIs formed from silhouettes where one leg/hand is behind the other leg/body. The weights are learned using the adaptive pooling technique mentioned in [24].

1-D Conv Pooling Since the output of Single Silhouette GEI generator is a sequence of single channel GEIs, these GEIs can be stacked together along the channel dimension. Once this is done, we can apply 1-D convolution [47] on this stacked pile of GEIs to get a single final GEI. The intuition behind using this approach is the same as that of weighted pooling.

Conv-GRU We used GRU to get a better fused representation of the intermediate GEIs. Since the dimensions of the input GEIs is large, using the standard GRU would have increased the network parameters drastically. We, instead, used Conv-GRU [3] which is more suitable for high dimensional image inputs. The equations of Conv-GRU, as shown below, are exactly same as that of standard GRU, except that the point-wise multiplication operation is replaced by convolution operation.

$$z_t^l = \sigma(W_z^l * x_t^l + U_z^l * h_{t-1}^l) \tag{3}$$

$$r_t^l = \sigma(W_r^l * x_t^l + U_r^l * h_{t-1}^l) \tag{4}$$

$$\tilde{h}_t^l = \tanh(W^l * x_t^l + U^l * (r_t^l \odot h_{t-1}^l)) \tag{5}$$

$$h_t^l = (1 - z_t^l)h_{t-1}^l + z_t^l \tilde{h}_t^l \tag{6}$$

Here, t represents the time-step, l represents the feature map number along the depth dimension, σ represents the sigmoid activation function, $*$ represents convolution operation and $\odot$ represents point-wise multiplication. z_t^l represents update gate and r_t^l represents the reset gate. W_z^l, U_z^l, W^l, U^l are the model parameters. h_t^l represents the hidden state at time-step t and l^{th} feature map.

3.2 Handling view variance

In this section we propose our model to handle view variance problem in gait recognition. This model takes GEI as input. The GEI of the same person may appear differently in different views. To overcome this challenge, one approach is to transform the probe and gallery

GEIs to one common view before comparing them, and in most cases, this common view is chosen as either of the probe and gallery views. However, the transformations performed by the current state-of-the-art methods are sub-optimal.

The current state-of-the-art models either transform all the gallery GEIs and probe/query GEIs to view angle of 90° [53], or transform the probe/query GEIs to gallery view [17] and then compare the GEIs. However, if the probe GEI has a view angle of 0° and the gallery GEI has view angle of 180°, then transforming the probe GEI from 0° view to 180° view may not be the right choice as shown in Fig. 4a. One may argue that in such cases we can transform both the gallery GEI and probe GEI to 90° view. However, consider the case when probe and gallery GEI both have a view angle of 0°. In such cases, transforming both the GEIs to 90° view may again be a sub-optimal choice as shown in Fig. 4b. The current state-of-the-art methods do not consider this issue.

In this work, we propose a novel method to overcome this shortcoming in current literature. Our model transforms a GEI from a source view to any target view (the target view can be sampled from a predefined finite set of views). This model is built around a Generative Adversarial Network (GAN) [14] framework, the details of which are described in the next subsection.

Given the shortcoming of the existing way of selecting the target view, we also present how to estimate the optimal view to transform the pair of GEIs, given the probe and gallery view angles. We provide details on this "Optimal View Predictor" in Section 3.2.2. Finally, we describe how the overall view-invariant gait-id model can be posed as a joint estimation of the view transformation, GAN as well as the optimal view predictor model parameters.

3.2.1 Network architecture

Our initial network consists mainly of 4 main parts: (1) Encoder, (2) View Classifier, (3) View Transformation Layer, and (4) GAN. The network details are shown in Fig. 5. The **Encoder** is used to obtain a low dimensional feature representation of the input GEIs. The **View Classifier** takes encoder's output as input and predicts the view angle of the GEI. The **View Transformation Layer** takes two inputs: encoder's output and an encoded vector representing transformation from initial view to target view. It transforms the feature representation from the initial view to the target view. The **GAN** takes this transformed feature representation and generates GEI in the target view. The encoded vector given to the View Transformation Layer is a vector containing values 1, −1, or 0. If we want a transformation from the i^{th} view to the j^{th} view ($j > i$), then the entries from i^{th} to j^{th} index in the transformation vector are set to 1 and rest are 0. −1 is used when $j < i$.

The loss functions are given in (7), (8) and (9). G represents generator, D represents the discriminator, E represents the encoder and V represents the view transformation layer. v represents the target view and u represents the initial view. x^k represents GEI in view k. $G(V(E(x^u), v, u)))$ represents the transformed GEI in view angle v from an initial GEI x^u in view angle u. L_{gan} is the GAN [14] loss, L_p is the pixel-wise loss and L_{total} is the combination of L_{gan} and L_p. γ is trade-off parameter and is set to 0.01. $\mathbb{E}$ is expectation. The view classifier is trained on cross entropy loss.

$$L_{gan} = \mathbb{E}[\log(D(x^v))] + \mathbb{E}[\log(1 - D(G(V(E(x^u), v, u))))] \tag{7}$$

$$L_p = \mathbb{E}\|G(V(E(x^u), v, u)) - x^v\|_1 \tag{8}$$

$$\min_{G,E,V} \max_{D} L_{total} = L_p + \gamma L_{gan} \tag{9}$$

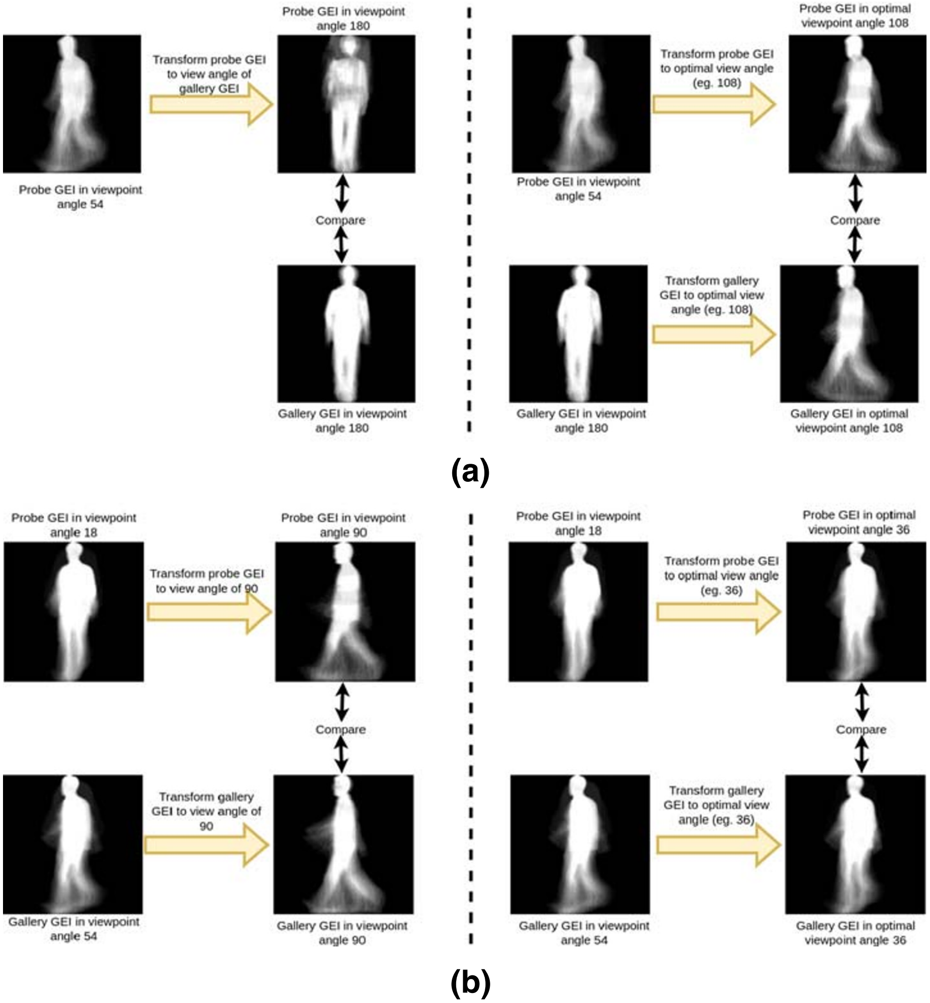
Probe GEI in viewpoint angle 54
Transform probe GEI to view angle of gallery GEI
Probe GEI in viewpoint angle 180
Compare
Gallery GEI in viewpoint angle 180
Probe GEI in viewpoint angle 54
Transform probe GEI to optimal view angle (eg. 108)
Probe GEI in optimal viewpoint angle 108
Compare
Gallery GEI in viewpoint angle 180
Transform gallery GEI to optimal view angle (eg. 108)
Gallery GEI in optimal viewpoint angle 108
(a)
Probe GEI in viewpoint angle 18
Transform probe GEI to view angle of 90
Probe GEI in viewpoint angle 90
Compare
Gallery GEI in viewpoint angle 54
Transform gallery GEI to view angle of 90
Gallery GEI in viewpoint angle 90
Probe GEI in viewpoint angle 18
Transform probe GEI to optimal view angle (eg. 36)
Probe GEI in optimal viewpoint angle 36
Compare
Gallery GEI in viewpoint angle 54
Transform gallery GEI to optimal view angle (eg. 36)
Gallery GEI in optimal viewpoint angle 36
(b)

Fig. 4 Issues with current state-of-the-art view invariant gait recognition methods. **a** Transforming from probe to gallery view (*left*) v/s Transforming to optimal view (*right*): In case when the probe and gallery views are far apart, transforming probe GEI to gallery view is not the optimal choice as the appearance of the GEI in the two views is very different and chances of erroneous transformation is high. Transforming to an optimal angle in between the probe and gallery view is a better choice. **b** Transforming to a pre-fixed view (90° in this case) (*left*) v/s Transforming to optimal view (*right*): In cases when the probe and gallery view angles are not far apart, transforming to a pre-fixed view, which may be very different from the probe and gallery views, is not an optimal choice. Transforming to a nearby angle is a better option

This network is similar to [17]. While a multi-task GAN is used in [17], we use a simple GAN [14] to perform our task of generating GEIs. The input to their network is a multi-channel Period Energy Image(PEI) whereas we give a single channel GEI as input.

3.2.2 Learning an optimal view

Given a pair of probe and gallery GEIs in different view angles, we propose to learn an optimal view which is most suited for comparing the two GEIs. In order to do this, we

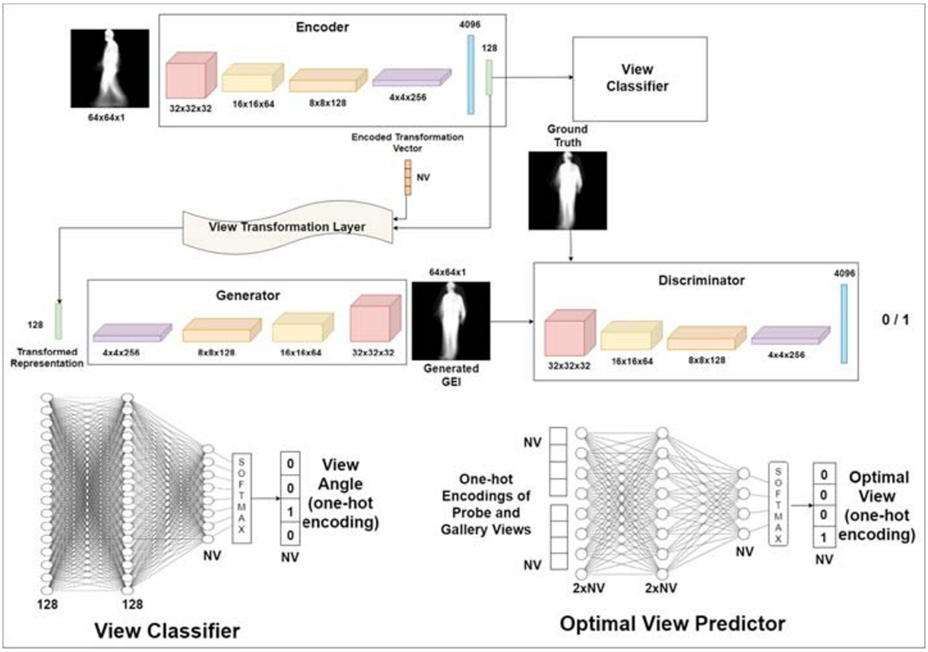

Fig. 5 View Invariant model: This model takes a GEI in a particular angle as input and transforms it to a GEI in a target angle. The encoder generates a low dimensional representation of the input GEI. The view transformation layer transforms the encoded representation to one in the target view. NV is the number of different view angles in the dataset. The generator generates GEIs in the target view. The target view is given by the optimal view predictor which takes one hot encodings (of size NV) of probe and gallery views given by the view classifier

add another module to our network, called **Optimal View Predictor** which takes one-hot encodings of the view angles of probe and gallery GEI and gives the corresponding optimal view. After that we transform the probe and gallery GEIs to the optimal view for comparison. The optimal view predictor is trained using the loss L_{pred} given in (10), (11) and (12). P represents the optimal view predictor, $O(v)$ is the one-hot encoding of view angle v and v_{opt} represents the optimal view given by P. x^v represents GEI in view v whereas $G(V(E(x^v), v_{opt}, v))$ represents the transformed GEI in view v_{opt}. Both x^v and x^u are the GEIs of the same person.

$$v_{opt} = P(O(v), O(u)) \tag{10}$$

$$d = \|G(V(E(x^v), v_{opt}, v)) - G(V(E(x^u), v_{opt}, u))\|_2 \tag{11}$$

$$\min_P L_{pred} = \mathbb{E}[d] \tag{12}$$

3.2.3 Joint estimation of parameters of *E*, *V*, *G*, *D* and *P*

As can be seen, the set of (7), (8), (9) and (10), (11), (12) are dependent on each other. We estimate these parameters in an iterative manner i.e. we first optimize over the parameters of G, V, E and D and then over the parameters of P. Hence, the (7) and (8) can be rewritten in terms of v_{opt}:

$$\begin{aligned} L_{gan} = & \ \mathbb{E}[\log(D(x^{v_{opt}}))] + \mathbb{E}[\log(1 - D(G(V(E(x^u), v_{opt}, u))))] \\ & + \mathbb{E}[\log(1 - D(G(V(E(x^v), v_{opt}, v))))] \end{aligned} \tag{13}$$

$$L_p = \mathbb{E}\|G(V(E(x^u), v_{opt}, u)) - x^v_{opt}\|_1 + \mathbb{E}\|G(V(E(x^v), v_{opt}, v)) - x^v_{opt}\|_1 \tag{14}$$

3.2.4 Ablation experiments

Using triplet loss to train the optimal view predictor We train the optimal view predictor using the triplet loss [18] instead of the fore-mentioned loss L_{pred}. The triplet loss $L_{triplet}$ is calculated as given in (15) to (18). P represents the optimal view predictor, $O(v)$ is the one-hot encoding of view angle v and v_{opt} represents the optimal view given by P. $x^v_{positive}$ represents the positive example GEI, $x^v_{negative}$ represents the negative example GEI and x^u_{anchor} represents the anchor GEI. $x^v_{positive}$ and x^u_{anchor} are the GEIs of the same person whereas $x^v_{negative}$ is the GEI of another person. x^v_k represents GEI in view v whereas $G(V(E(x^v_k), v_{opt}, v))$ represents the transformed GEI in view v_{opt} where $k \in \{positive, negative, anchor\}$. m is the margin which is set to 1.0.

$$v_{opt} = P(O(v), O(u)) \tag{15}$$

$$d_1 = \mathbb{E}\|G(V(E(x^v_{positive}), v_{opt}, v)) - G(V(E(x^u_{anchor}), v_{opt}, u))\|_2 \tag{16}$$

$$d_2 = \mathbb{E}\|G(V(E(x^v_{negative}), v_{opt}, v)) - G(V(E(x^u_{anchor}), v_{opt}, u))\|_2 \tag{17}$$

$$\min_P L_{triplet} = max(0, d_1 - d_2 + m) \tag{18}$$

Thus, triplet loss [18] tries to bring the transformed GEIs in view angle v_{opt} of the same person closer whereas the GEI of different person is moved further away. Using triplet loss we got competitive results but the best results (as shown in the Section 5.2.2) were obtained using L_{pred} loss to train the optimal view predictor.

Using GEI information in addition to View information to predict optimal view As can be seen in (10), we are only using the view angles of the probe and gallery GEIs to predict the optimal view. In (19), we incorporate GEI information along with the view to predict the optimal view. This is done by appending the output vector from the encoder to the one-hot vector representing the view and passing the vector obtained after appending to the Optimal View Predictor Module as shown in the following equation.

$$v_{opt} = P(O(v)^\frown E(x^v_{probe}), O(u)^\frown E(x^u_{gallery})) \tag{19}$$

In (19), P represents the optimal view predictor and E represents the encoder. $O(v)$ is the one-hot encoding of view angle v and v_{opt} represents the optimal view given by P. x^v_{probe} represents the probe GEI in view v and $x^u_{gallery}$ represents gallery GEI in view u. $E(x^v_{probe})$ and $E(x^u_{gallery})$ represents the output of encoder for GEIs x^v_{probe} and $x^u_{gallery}$ respectively. $\frown$ represents the concatenation (of vectors) operation.

4 Experiments details

4.1 Datasets

CASIA-B Yu et al. [55] dataset is a large multi-view dataset of 124 subjects. The dataset contains sequences captured from 11 different view angles ranging from 0° to 180° with an interval of 18°. The dataset has 6 sequences under normal conditions (named ‘nm-01’ to ‘nm-06’), 2 sequences of walking in a coat (named ‘cl-01’ and ‘cl-02’) and 2 sequences of walking with a bag (named ‘bg-01’ and ‘bg-02’) for each of the subjects.

The dimension of each silhouette image is 320×240. Each sequence is 2 to 3 gait cycles long. Out of the 124 subjects, 93 are males and 31 are females. Most of the subjects are aged between 20 and 30 years. There are 10 silhouette sequence corresponding to each person in a particular view -6 normal $+2$ with coat $+2$ with a bag. Since there are 11 different views, so the total number of sequences in the dataset are $11 \times 10 \times 124 = 13640$. We have shown results on only the sequences under normal clothing and carrying conditions.

OU-ISIR Large Population gait dataset [23] is a large dataset containing over 4000 subjects of different ages, ranging from 1 year to 94 years. We have used the latest version of the dataset, Version 2, which has 4016 subjects. These subjects are captured from four different view angles $-55°$, 65°, 75° and 85°. The dataset is divided further into two subsets-set A and set B. We have used set A which has two sequences per identity per view angle.

The major strength of this dataset is the large number of subjects it contains as opposed to other datasets. In addition to this, the dataset is more gender balanced with nearly equal number of males and females, and the age range of the subjects is also much wider compared to other datasets. The silhouette sequences are of much better quality since they are manually checked. In both the datasets, the length of a gait cycle is 25–30 frames [2].

4.2 GEI construction model

For our experiments to handle the missing data, we take the first 62 subjects of the CASIA-B [55] dataset in the training set and the remaining 62 subjects in the test set. In the test set, sequences ‘nm-05’ and ‘nm-06’ form the probe set and the sequences ‘nm-01’–‘nm-04’ form the gallery set. For the OU-ISIR [23] gait dataset, we follow the same train-test split as followed by ITCNet [2]. We select 3254 subjects and the randomly pick 2254 subjects in the training set, 500 subjects in the validation set and 500 subjects in the test set and perform 5 fold cross validation. To train our model we use ADAM optimizer and the learning rate is set to 0.8×10^{-4}. The batch size while training is 120.

The silhouette sequences in the CASIA-B dataset capture the entire scene in the image where the subject (white image in black background) is only a part of the scene and may be located at different locations in the image/silhouette. While giving input to the network, the subject is cropped by drawing a bounding box around it and then the cropped image is resized to dimension of 64×64. In case of weighted pooling, 1-D conv pooling and Conv-GRU the length of input sequence is fixed to 30 as this is the typical length of a gait cycle in these datasets whereas the average pooling method is independent of sequence length.

4.3 View invariant model

In our experiments to handle view variance, we take the first 62 subjects of CASIA-B [55] dataset in train set and the next 62 subjects in test set. In the test set, the sequences ‘nm-05’ and ‘nm-06’ form the probe set and the sequences ‘nm-01’–‘nm-04’ form the gallery set. The input to the view invariant model is a GEI. In the case of CASIA-B dataset, GEIs corresponding to each sequence are provided whereas in the case of OU-ISIR dataset [23] we construct GEIs by averaging the silhouettes of a sequence. For OU-ISIR dataset [23], we divide the dataset into 5 equal parts and perform 5 fold cross validation. This is done to ensure fair comparison with other state-of-the-art methods. We use ADAM optimizer to train the model. The learning rate while training the GAN, encoder and view transform layer is 10^{-6} and while training the optimal view predictor is 10^{-7}. The batch size while training is 500.

As mentioned before, we perform the training of the view invariant module in multiple phases. In the first phase, we only train the encoder, GAN and view transformation layer. Then we train the view classifier using cross-entropy loss. In the second phase, we train the optimal view predictor and fine-tune the entire module.

4.4 Evaluation metric

We use Rank-1 identification accuracy(%) as our evaluation metric. During the test time, euclidean distance is used to find the nearest example to the probe in the gallery set.

5 Results

5.1 Handling missing data

We compare the results of our GEI construction model with ITCNet [2] which, to the best of our knowledge, is the only work that handles the problem of incomplete silhouette sequences. We compare the results on both the datasets. The results are presented for the view angle of 90° in case of CASIA-B [55] in Table 1 and view angle of 85° in case of OU-ISIR dataset [23] in Table 2, since ITCNet [2] presents results only on these angles. As can be seen, we are outperforming ITCNet in all cases on CASIA-B and most of the cases on OU-ISIR. In Tables 1 and 2, the best accuracies reported (by either [2] or our proposed model) for each of the different sequence lengths are highlighted in bold.

5.1.1 GEI pooling approaches beyond average pooling

In this section we present the results of applying pooling techniques such as weighted pooling, 1-D conv pooling and Conv-GRU on the intermediate GEIs generated by the single

Table 1 Rank-1 identification accuracy(%) of GEI construction module on CASIA-B dataset(view angle of 90°)

#Silhouettes	ITCNet [2]	Our model
1	50.01	**57.6**
2	50.09	**62.7**
4	60.02	**77.11**
6	75.10	**81.35**
8	80.00	**83.89**
10	80.06	**86.44**
13	77.12	**88.98**
15	85.24	**88.98**
27	85.30	**92.37**

Table 2 Rank-1 identification accuracy(%) of GEI construction module on OU-ISIR dataset(view angle of 85°)

#Silhouettes	ITCNet [2]	Our model
1	53.40	**62.40**
3	65.11	**67.40**
5	71.19	**72.00**
8	**77.53**	77.20
10	**80.46**	79.60
13	82.59	**83.20**
15	84.22	**85.60**
18	85.76	**86.80**
20	86.00	**86.80**

silhouette GEI generator. The results are shown for view angle of 90° in case of CASIA-B dataset in Table 3.

5.1.2 Performance of GEI construction module for different view angles

In order to see the performance of average pooling method for different view angles, we train different networks for different view angles and show the results in Table 4. In this table, the probe and gallery sequences are captured from the same views in each case. As can be seen in the table, our GEI reconstruction module performs well across different view angles.

5.2 Handling view variance

In this section, we present the results of our view invariant model in Table 5, Table 6, Fig. 6 and Fig. 7. The results on CASIA-B dataset are presented in Fig. 6 and Table 6 whereas the results on OU-ISIR dataset are presented in Fig. 7. In Fig. 6, we compare our model with SPAE [54] and GaitGAN [53], both of which take GEIs as input. The same evaluation protocol is followed as in these works. The training set contains the first 62 subjects and the test set contains next 62 subjects. In Fig. 7, we compare our model's performance on the OU-ISIR dataset with other state of the art methods. As can be seen in Fig. 7, we exclude the cases when the probe view and gallery view are the same.

Table 3 Performance of other pooling methods : Rank-1 accuracy (%) on CASIA-B dataset for view angle of 90°

#Silhouettes	Conv-GRU	Weighted pooling	1-D conv pooling
1	16.81	20.17	15.12
2	26.89	28.57	32.77
4	47.9	48.74	50.42
6	61.34	66.39	63.03
8	65.55	72.27	68.91
10	78.99	74.79	77.31
13	82.35	81.51	78.15
15	84.03	84.03	82.35
20	88.24	89.91	89.07

These methods perform well when the number of available silhouettes are 10 or above. However, when the number of available silhouettes are less than 10, average pooling gives better results

Table 4 Rank 1 accuracy(%) of GEI construction module on different view angles

N	View angle									
	0°	18°	36°	54°	72°	108°	126°	144°	162°	180°
1	49.57	41.66	38.33	45.83	49.83	50	41.67	54.16	43.33	55
2	64.7	61.66	56.66	73.33	55	66.66	55.83	60.83	64.16	65.83
4	82.35	77.5	80	80	74.16	74.16	71.66	74.16	70.83	82.5
6	84.03	85	89.16	83.33	78.33	77.5	80	81.67	79.16	86.67
8	90.75	87.5	93.33	87.5	79.16	76.67	83.33	85	79.16	91.67
10	90.75	87.5	93.33	85.83	80.83	80	85	85.83	82.5	91.67
13	94.95	89.16	97.5	87.5	84.16	82.5	87.5	90.83	87.5	94.16
15	95.95	90.83	95.83	90	85	82.5	88.33	90.83	84.16	94.16
20	95.79	93.3	97.5	94.16	86.66	84.16	89.16	90.83	89.16	95.83
30	94.95	97.5	97.5	95	87.5	87.5	89.16	91.67	92.5	95.83

N is the number of available silhouettes

In addition, we compare the results of our method with [48] on the OU-ISIR dataset in Table 5. In [48], Takemura et al. propose four different architectures: a CNN with contrastive loss function (i.e., named *2in* in this paper), a CNN with triplet ranking loss (i.e., named *3in* in this paper), a CNN with low-level difference structure (i.e, named *diff* in the paper) and a CNN with both low and high level difference structure (i.e., named *2diff* in the paper). We get better results than them when compared to single network architectures proposed in their work. Only when multiple networks proposed in [48] are combined together, they get a better performance than our method. As stated in their work, their *2in* and *3in* were better than *diff* and *2diff* for cases with large angular differences, and the opposite result was obtained for cases with small angular differences. However, we propose a single network architecture to handle all these cases and this gives us an advantage over them. Also, our

Table 5 Comparison of rank-1 accuracies(%) of our model with [48] on OU-ISIR dataset

	Angular difference				
	0	10	20	30	mean
2in	97.9	97.6	95.6	92.0	96.5
3in	98.5	98.2	96.4	92.3	97.1
diff	98.7	98.5	97.2	94.7	97.7
2diff	99.1	99.0	98.0	95.1	98.3
2in+diff	99.3	99.2	98.6	96.9	98.8
3in+2diff	99.2	99.2	98.6	97.0	98.8
Ours	98.8	98.7	98.12	96.73	98.32

The results in the last column are the average across all probe and gallery view angles. Our method performs better than the individual networks proposed in [48]. Only when multiple networks proposed in [48] are combined together, they get a better performance than our method. Also, as stated in their work, their *2in* and *3in* were better than *diff* and *2diff* for cases with large angular differences, and the opposite result was obtained for cases with small angular differences. However, we propose a single network architecture to handle all these cases and this gives us an advantage over them

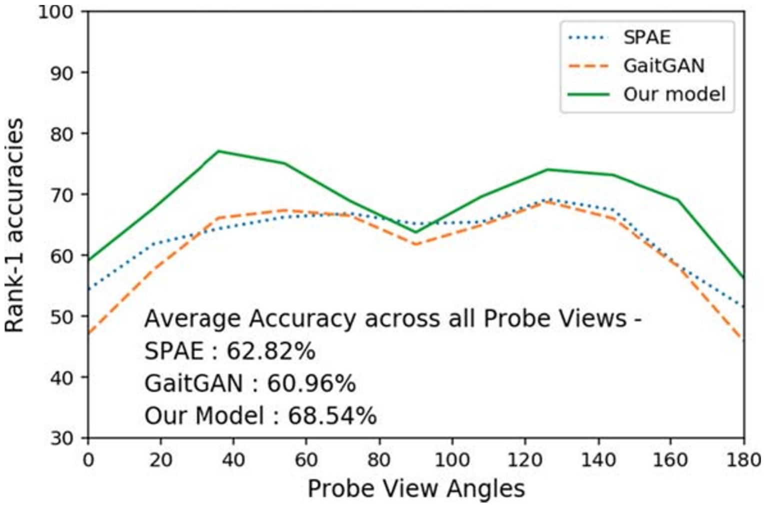

Fig. 6 Comparison of rank 1 accuracy(%) of our view invariant module with other GEI based models, SPAE [54] and GaitGAN [53], on CASIA-B dataset. The x-axis represents the probe view angle whereas the y-axis represents the average rank-1 accuracy across all gallery view angles

proposed network does not require combining two different network and works as a single end-to-end network

In Table 6, we compare our model's performance with other view invariant models that take different gait features other than GEIs as input such as Period Energy Images [17] or Mixture of GEIs [51] or silhouettes [10]. We observe that we perform better than most of the state-of-the art methods except [17] and [51] on CASIA-B dataset. However, our model performs better than them on the OU-ISIR dataset. We believe that using better features like PEIs [17] and mixture of GEIs [51] helps in improving the model performance. However, such features demand the availability of complete data which may not be possible in practical scenarios. For the results presented in Table 6, the model was trained with first 74 subjects in the train set and tested on the remaining 50 subjects of the CASIA-B dataset. This was done to ensure that the same protocol is followed as other works presented in Table 6.

5.2.1 Improvement using optimal view predictor

Figure 8 shows the advantage of using Optimal View Predictor on our initial model described in Section 3.2.1. We compare the results obtained by transforming from probe view to gallery view with the results obtained by using the Optimal View Predictor to obtain optimal view to transform the GEIs. We apply our method to [17], using GEIs as input, to show the effectiveness of transforming to an optimal view, and get an improvement in average rank-1 accuracy from 70.1% to 71.28% as shown in Table 7.

5.2.2 Ablation studies

In this section, we present results obtained by performing ablation on loss function and input to optimal view predictor. In the first first experiment, we use $L_{triplet}$ loss defined in (15)–(18) to train the optimal view predictor instead of the L_{pred} loss defined in (10)–(12). Apart from this we also perform an experiment where we give the encoder's output to the

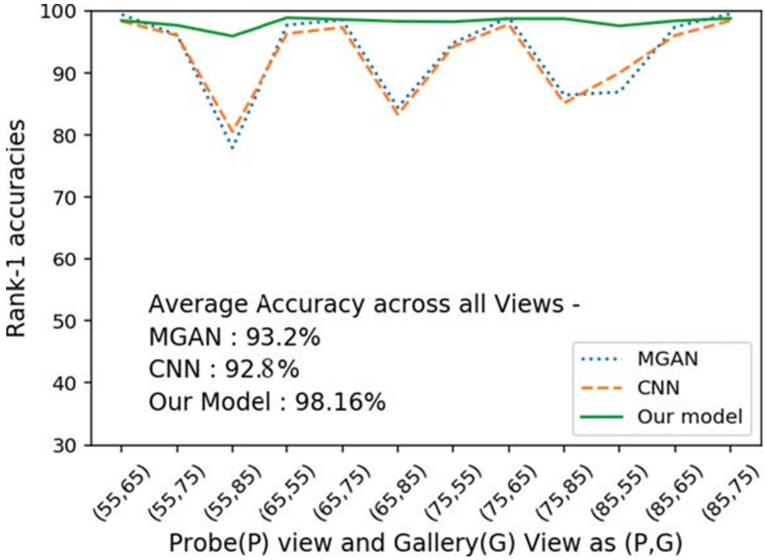

Fig. 7 Comparison of rank 1 accuracy(%) of our view invariant module with other state-of-the-art models, MGAN [17] and CNN [51], on OU-ISIR dataset. The OU-ISIR dataset contains sequences captured only in 4 different view angles as shown in the figure. We do not consider the cases when probe view and gallery view are the same

optimal view predictor along with the view angle information for each GEI. The results of these experiments are presented in Table 8.

5.3 An end-to-end pipeline towards gait based recognition on RGB input sequence

Our method, which comprises of the GEI Construction Module and the View Invariant Module, takes silhouette sequence as inputs. In order to show the real-world usability of the proposed gait-based identification framework, we perform an end-to-end experiment. For this experiment, we assume that only RGB image sequences/videos are available as inputs

Table 6 Comparison of rank-1 accuracies(%) of our model with other state-of-the-art models which use different gait features other than GEIs

	54°	90°	126°	Mean
C3A [52]	75.7	63.7	74.8	71.4
ViDP [21]	64.2	60.4	65	63.2
VTMSVR [25]	55	46	54	51
MGAN [17]	84.2	72.3	83	79.8
CNN [51]	94.6	88.3	93.8	92.2
GaitSet [10]	96.9	91.7	97.8	95.46
Ours	79.3	68.9	78.2	75.46

Results are shown for probe view angles 54°, 90°, 126°, excluding identical view cases (i.e. for each probe view angle, an average of rank-1 accuracies are taken over all the gallery views except those views which are same as the probe view). All the models of this table are trained on the first 74 subjects of CASIA-B dataset

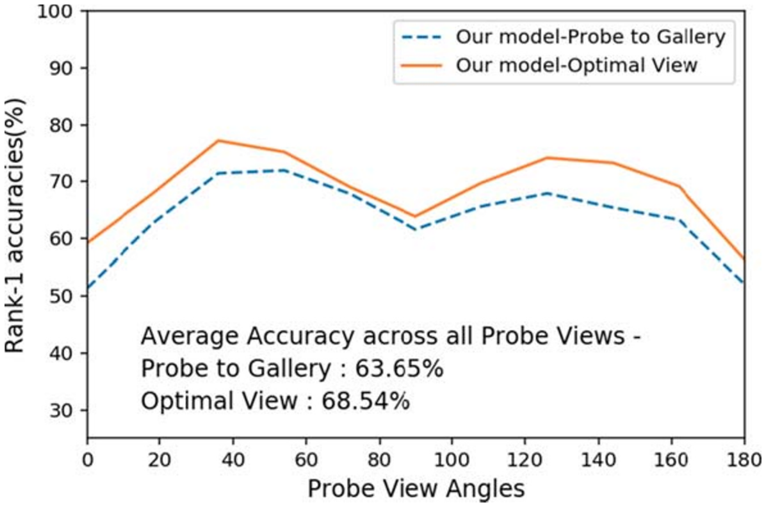

Fig. 8 Improvement in Rank-1 accuracy on using the Optimal View Predictor instead of probe to gallery view transformation on CASIA-B dataset

to the framework instead of the silhouette sequence. Now, as a first step in the end-to-end pipeline, we employed a Mask R-CNN [16] network, pre-trained on COCO dataset [30], to extract silhouette sequence from the RGB video. The reason we extract silhouettes from the input RGB images is because the gait-based methods, unlike person re-identification techniques, are not designed to use any chromatic appearance information contained in the RGB images/videos. Due to unavailability of RGB video gait dataset, we have used a subset of publicly available person re-identification RGB dataset, PRID-2011 [19] in this experiment.

We show results on sequences of PRID-2011 dataset which were captured at view angle of 90°. Additionally, since PRID-2011 [19] is a person re-identification dataset, it contains sequences that have self-occlusion (which partially covers the human body shape of a

Table 7 Improvement in Rank-1 accuracy of MGAN [17] on using the Optimal View Predictor instead of probe to gallery view transformation on CASIA-B dataset

Probe Angle	MGAN [17] Probe to gallery	MGAN Optimal view (ours)
0°	58	63.19
18°	65.9	72.06
36°	73	77.12
54°	77.1	76.02
72°	73.1	70.38
90°	67.3	66.13
108°	69.9	73.38
126°	78.6	77.93
144°	77.9	76.09
162°	71.9	70.45
180°	58.4	61.36
Average	70.1	71.28

The models are trained on the first 62 subjects and tested on the remaining 62 subjects

Table 8 Ablation Results - Rank 1 accuracies(%) obtained by using triplet loss to train the optimal view predictor and concatenating encoder output to view angle while giving input to the optimal view predictor

Probe Angle	Using triplet loss to train optimal view predictor	Encoder output + view angle as input to optimal view predictor
0°	58.13	56.16
18°	68.25	66.28
36°	71.77	73.75
54°	73.38	72.80
72°	69.87	68.03
90°	65.98	63.12
108°	69.65	68.91
126°	72.36	70.53
144°	73.75	71.85
162°	68.91	69.13
180°	56.67	57.69
Average	68.07	67.12

The models are trained on the first 62 subjects of CASIA-B dataset and tested on the remaining 62 subjects

subject). As the re-identification methods use chromatic appearance information, they are relatively less affected by self-occlusions. However, such observations with self-occlusions massively degrade the performance of gait based recognition systems that do not use any chromatic information and work with minimal input information of the human body shape. Hence, typical gait based approaches use sequences which contain images of full human body shape. To ensure this, we choose 32 identities from the PRID-2011 dataset, that satisfy this condition. Out of the 32 identities, 8 identities were used in the train set to fine-tune our network and the remaining 24 identities were in the test set. We performed 4-fold cross validation on the 32 identities.

Table 9 shows the results of the GEI construction module on this dataset. It can be observed from Table 9 that the proposed end-to-end framework achieves a very respectable performance in terms of rank-1 accuracy even when it is applied in a more real-world scenario with RGB image sequences as inputs. We present more details on the dataset creation in the Supplementary Material provided along with the paper. This experiment thus demonstrates the ability of our method to work in an end-to-end manner along with an off-the-shelf gait segmentation module.

Table 9 Rank-1 identification accuracy(%) of GEI construction module on 32-ids subset of PRID-2011 dataset(view angle of 90°)

#Silhouettes	Rank-1 Accuracy(%)
1	33.3
3	54.16
5	58.3
8	70.8
10	75.0
13	79.1
15	79.1
18	83.3
20	87.5

Table 10 Performance of the combined model

	Number of Silhouettes			
	3	8	20	40
Baseline	7.53	21.69	36.39	48.38
Combined Model	18.08	42.64	46.20	60.33

The models are trained on the first 62 subjects of the CASIA-B dataset and tested on the remaining 62 subjects. The results show that our model is robust to the challenges posed by missing data/incomplete sequences captured in different view angles making it more suitable for practical scenarios

5.4 Combining the two models

In this section we combine the GEI construction module and view invariant module and show the advantage of using the GEI construction module in cases when the number of silhouette frame are less. The input here are incomplete probe and gallery sequences of silhouettes which are classified into one of the possible view using a view classifier. The sequences are then given to the GEI construction network for that particular view. The GEI construction module generates probe and gallery GEIs corresponding to the input sequence in their respective view angles. Now, the probe and gallery GEIs are given to the view invariant module, where they are compared for the task of identification. We compare our combined model performance in Table 10 with the baseline case when the available silhouettes are simply averaged to create a GEI(which is the general straight-forward method to construct a GEI). This GEI is then given to the view invariant module.

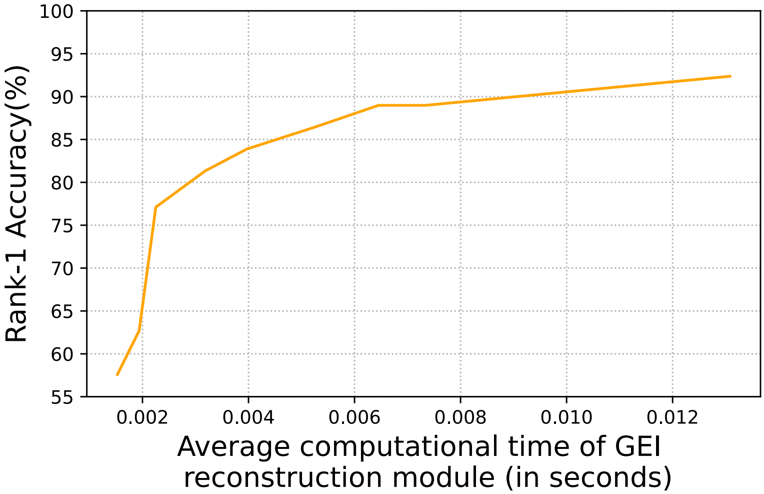

Fig. 9 Computational time v/s Accuracy comparison for the GEI construction module. As the number of silhouettes in the input sequence increases, the computational time and identification accuracy also increases. However, after the length of input sequences approaches the length of gait cycle, accuracy remains almost same. The above results are shown for CASIA-B dataset

5.5 Computational time complexity of our model

In this section, we discuss the computational time complexity of different modules of our model. The GEI construction module produces a GEI corresponding to each silhouette in the input sequence, and finally combines them to form the final GEI. Therefore, the time complexity of the GEI construction module is $\mathcal{O}(l)$ where l is the length of the input sequence available for a particular identity/person. The view invariant model takes the GEI as an input and compares it with the GEI representation of every other identity in the gallery set, after transforming them to a common optimal view. However, the computational time complexity is same of each identity and is dependent on the size of the gallery set. Therefore, the time complexity of the view invariant module can be represented as $\mathcal{O}(D)$, where D is the number of identities in the gallery set. In Fig. 9, we compare computational time with accuracy of the GEI construction module. The experiment was run on a system with Intel(R) i7-8700 CPU processor with 32 GB RAM, 1080Ti GPU.

6 Conclusion

In this work, we presented a novel and robust approach to perform the task of gait based person recognition under real-life, practical scenarios. We presented methods to handle the challenges posed by incomplete silhouette sequences and multi-view sequences. We, first, looked into the incomplete sequence problem. We proposed a method which took incomplete silhouette sequences as input and constructed accurate GEIs, similar to the GEIs constructed from complete gait cycle sequence (if it were available). We used a Single Silhouette GEI Generator which generates GEIs, from single silhouettes, which are then fused together to get a final GEI. We explored and compared different techniques to fuse the generated GEIs. We then addressed the issue of view variance. We proposed a novel method to jointly, estimate the optimal view for comparison of the probe and gallery observations depending on their respective initial view angles, as well as learn the transformations between GEIs across these views. We showed the advantages of using such an optimal transformation, both qualitatively and quantitatively. Through extensive experiments on two gait benchmark datasets, we showed that our proposed framework outperforms most of the state-of-the-art methods.

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/s11042-020-10132-z).

Funding This work is partially supported by 1. Pratiksha Trust, Bangalore and 2. Start-up Research Grant (SRG) from SERB, DST, India to Anirban Chakraborty (Project file number: SRG/2019/001938).

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

References

1. Ariyanto G, Nixon MS (2011) Model-based 3d gait biometrics. In: 2011 International joint conference on biometrics (IJCB), IEEE, pp 1–7

2. Babaee M, Li L, Rigoll G (2019) Person identification from partial gait cycle using fully convolutional neural networks. Neurocomputing 338:116–125
3. Ballas N, Yao L, Pal C, Courville A (2015) Delving deeper into convolutional networks for learning video representations. arXiv:1511.06432
4. Bashir K, Xiang T, Gong S (2009) 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009). In: Gait recognition using Gait Entropy Image, pp 1–6. https://doi.org/10.1049/ic.2009.0230
5. Ben X, Meng W, Yan R, Wang K (2012) An improved biometrics technique based on metric learning approach. Neurocomputing 97:44–51
6. Bobick AF, Johnson AY (2001) Gait recognition using static, activity-specific parameters. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, IEEE, vol 1, pp I–I
7. Bodor R, Drenner A, Fehr D, Masoud O, Papanikolopoulos N (2009) View-independent human motion classification using image-based reconstruction. Image Vis Comput 27(8):1194–1206
8. Bouchrika I, Nixon MS (2007) Model-based feature extraction for gait analysis and recognition. In: International conference on computer vision/computer graphics collaboration techniques and applications, Springer, pp 150–160
9. Boulgouris NV, Chi ZX (2007) Human gait recognition based on matching of body components. Pattern Recogn 40(6):1763–1770
10. Chao H, He Y, Zhang J, Feng J (2019) Gaitset: regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8126–8133
11. De Marsico M, Mecca A, Barra S (2019) Walking in a smart city: investigating the gait stabilization effect for biometric recognition via wearable sensors. Comput Electr Eng 80:106501
12. Goffredo M, Bouchrika I, Carter JN, Nixon MS (2010) Self-calibrating view-invariant gait biometrics. IEEE Trans Syst Man Cybern B (Cybern) 40(4):997–1008
13. González I, López-Nava IH, Fontecha J, Muñoz-Meléndez A, Pérez-SanPablo AI, Quiñones-Urióstegui I (2016) Comparison between passive vision-based system and a wearable inertial-based system for estimating temporal gait parameters related to the gaitrite electronic walkway. J Biomed Inform 62:210–223
14. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
15. Han J, Bhanu B (2006) Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 28(2):316–322
16. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
17. He Y, Zhang J, Shan H, Wang L (2019) Multi-task gans for view-specific feature learning in gait recognition. IEEE Transactions on Information Forensics and Security 14(1):102–113
18. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:1703.07737
19. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on image analysis, Springer, pp 91–102
20. Hossain MA, Makihara Y, Wang J, Yagi Y (2010) Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. Pattern Recogn 43(6):2281–2291
21. Hu M, Wang Y, Zhang Z, Little JJ, Huang D (2013) View-invariant discriminative projection for multi-view gait-based human identification. IEEE Transactions on Information Forensics and Security 8(12):2034–2045
22. Huang GB, Zhu QY, Siew CK (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE international joint conference on neural networks (IEEE Cat. no. 04CH37541), IEEE, vol 2, pp 985–990
23. Iwama H, Okumura M, Makihara Y, Yagi Y (2012) The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Transactions on Information Forensics and Security 7(5):1511–1521
24. Kar A, Rai N, Sikka K, Sharma G (2017) Adascan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3376–3385
25. Kusakunniran W, Wu Q, Zhang J, Li H (2010) Support vector regression for multi-view gait recognition based on local motion feature selection. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, pp 974–981

26. Kusakunniran W, Wu Q, Zhang J, Li H, Wang L (2014) Recognizing gaits across views through correlated motion co-clustering. IEEE Trans Image Process 23(2):696–709
27. Kusakunniran W, Wu Q, Zhang J, Ma Y, Li H (2013) A new view-invariant feature for cross-view gait recognition. IEEE Transactions on Information Forensics and Security 8(10):1642–1653
28. Kyrarini M, Wang X, Gräser A (2015) Comparison of vision-based and sensor-based systems for joint angle gait analysis. In: 2015 IEEE international symposium on medical measurements and applications (memea) proceedings, IEEE, pp 375–379
29. Lam TH, Cheung KH, Liu JN (2011) Gait flow image: a silhouette-based gait representation for human identification. Pattern Recogn 44(4):973–987
30. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P., Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision, Springer, pp 740–755
31. Liu Z, Sarkar S (2004) Simplest representation yet for gait recognition: averaged silhouette. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004., IEEE, vol 4, pp 211–214
32. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
33. López-Fernández D, Madrid-Cuevas FJ, Carmona-Poyato A, Marín-Jiménez MJ, Muñoz-Salinas R, Medina-Carnicer R (2016) Independent gait recognition through morphological descriptions of 3d human reconstructions. Image Vis Comput 48:1–13
34. Luo J, Tang J, Tjahjadi T, Xiao X (2016) Robust arbitrary view gait recognition based on parametric 3d human body reconstruction and virtual posture synthesis. Pattern Recogn 60:361–377
35. Luo J, Tjahjadi T (2020) Multi-set canonical correlation analysis for 3d abnormal gait behaviour recognition based on virtual sample generation. IEEE Access 8:32485–32501
36. Makihara Y, Sagawa R, Mukaigawa Y, Echigo T, Yagi Y (2006) Gait recognition using a view transformation model in the frequency domain. In: European conference on computer vision, Springer, pp 151–163
37. Marín-Jiménez MJ, Castro FM, Guil N, de la Torre F, Medina-Carnicer R (2017) Deep multi-task learning for gait-based biometrics. In: 2017 IEEE International conference on image processing (ICIP), IEEE, pp 106–110
38. Muramatsu D, Makihara Y, Yagi Y (2014) Are intermediate views beneficial for gait recognition using a view transformation model? In: Proc. of the 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV 2014), pp 222–227
39. Nandi GC, Semwal VB, Raj M, Jindal A (2016) Modeling bipedal locomotion trajectories using hybrid automata. In: 2016 IEEE Region 10 conference (TENCON), IEEE, pp 1013–1018
40. Ortells J, Mollineda RA, Mederos B, Martín-Félez R (2017) Gait recognition from corrupted silhouettes: a robust statistical approach. Mach Vis Appl 28(1-2):15–33
41. Patil P, Kumar KS, Gaud N, Semwal VB (2019) Clinical human gait classification: extreme learning machine approach. In: 2019 1st international conference on advances in science, engineering and robotics technology (ICASERT), IEEE, pp 1–6
42. Phinyomark A, Petri G, Ibáñez-Marcelo E, Osis ST, Ferber R (2018) Analysis of big data in gait biomechanics: current trends and future directions. J Med Biol Eng 38(2):244–260
43. Semwal VB, Gaud N, Nandi G (2019) Human gait state prediction using cellular automata and classification using elm. In: Machine intelligence and signal analysis, Springer, pp 135–145
44. Semwal VB, Kumar C, Mishra PK, Nandi GC (2016) Design of vector field for different subphases of gait and regeneration of gait pattern. IEEE Trans Autom Sci Eng 15(1):104–110
45. Shiraga K, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2016) Geinet: view-invariant gait recognition using a convolutional neural network. In: 2016 International conference on biometrics (ICB), IEEE, pp 1–8
46. Sokolova A, Konushin A (2017) Pose-based deep gait recognition. arXiv:1710.06512
47. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
48. Takemura N, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2017) On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Transactions on Circuits and Systems for Video Technology
49. Tang J, Luo J, Tjahjadi T, Gao Y (2014) 2.5 d multi-view gait recognition based on point cloud registration. Sensors 14(4):6124–6143
50. Wang C, Zhang J, Pu J, Yuan X, Wang L (2010) Chrono-gait image: a novel temporal template for gait recognition. In: European conference on computer vision, Springer, pp 257–270

51. Wu Z, Huang Y, Wang L, Wang X, Tan T (2017) A comprehensive study on cross-view gait based human identification with deep cnns. IEEE Trans Pattern Anal Mach Intell 39(2):209–226
52. Xing X, Wang K, Yan T, Lv Z (2016) Complete canonical correlation analysis with application to multi-view gait recognition. Pattern Recogn 50:107–117
53. Yu S, Chen H, Reyes G, Edel B, Poh N (2017) Gaitgan: invariant gait feature extraction using generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 30–37
54. Yu S, Chen H, Wang Q, Shen L, Huang Y (2017) Invariant feature extraction for gait recognition using only one uniform model. Neurocomputing 239:81–93
55. Yu S, Tan D, Tan T (2006) A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th international conference on pattern recognition (ICPR’06), IEEE, vol 4, pp 441–444
56. Zhao G, Liu G, Li H, Pietikainen M (2006) 3d gait recognition using multiple cameras. In: 7th international conference on automatic face and gesture recognition (FGR06), IEEE, pp 529–534

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Utkarsh Shreemali completed his M.Tech. from the Department of Computational and Data Sciences, Indian Institute of Science (IISc), Bangalore under the guidance of Prof. Anirban Chakraborty. His research interests include machine learning, computer vision and person re-identification. On completion of the M.Tech. program, Utkarsh has taken up employment at Qualcomm India Private Limited, Hyderabad, India.

Anirban Chakraborty received his Ph.D. in Electrical Engineering from the University of California, Riverside in 2014. Subsequently, he held research fellow positions with the National University of Singapore and Nanyang Technological University. After that, Anirban worked as a computer vision researcher at the Robert Bosch Research and Technology Centre, India. Currently, he is an assistant professor at the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore. His research interests lie in the broad areas of computer vision, machine learning, optimization etc. and their applications in problems such as data association over large graphs, data fusion, video surveillance problems, video-based biometrics, multimedia etc. He is also keen to explore how visual analytics can be utilized in answering some of the most fundamental questions in biology and healthcare.