A non-linear view transformations model for cross-view gait recognition

Article *in* Neurocomputing · April 2020

DOI: 10.1016/j.neucom.2020.03.101

Muhammad Hassan Khan
Universität Siegen
22 PUBLICATIONS 87 CITATIONS
SEE PROFILE

Muhammad Shahid Farid
University of the Punjab
44 PUBLICATIONS 225 CITATIONS
SEE PROFILE

Marcin Grzegorzek
Universität zu Lübeck
154 PUBLICATIONS 719 CITATIONS
SEE PROFILE

A non-linear view transformations model for cross-view gait recognition

Muhammad Hassan Khan[a,b,*], Muhammad Shahid Farid[b], Marcin Grzegorzek[a,c]

[a]*Research Group of Pattern Recognition, University of Siegen, Siegen, Germany.*
[b]*Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan.*
[c]*Institute of Medical Informatics, University of Lübeck, Lübeck, Germany.*

Abstract

Gait has emerged as an important biometric feature which is capable of identifying individuals at distance without requiring any interaction with the system. Various factors such as clothing, shoes, and walking surface can affect the performance of gait recognition. However, cross-view gait recognition is particularly challenging as the appearance of individual's walk drastically changes with the change in the viewpoint. In this paper, we present a novel view-invariant gait representation for cross-view gait recognition using the spatiotemporal motion characteristics of human walk. The proposed technique trains a deep fully connected neural network to transform the gait descriptors from multiple viewpoints to a single canonical view. It learns a single model for all the videos captured from different viewpoints and finds a shared high-level virtual path to project them on a single canonical view. The proposed deep neural network is learned only once using the spatiotemporal gait representation and applied to testing gait sequences to construct their view-invariant gait descriptors which are used for cross-view gait recognition. The experimental evaluation is carried out on two large benchmark cross-view gait datasets, CASIA-B and OU-ISIR large population, and the results are compared with current state-of-the-art methods. The results show that the proposed algorithm outperforms the state-of-the-art methods in cross-view gait recognition.

Keywords: Cross-view gait recognition, view transformations, Spatiotemporal features

1. Introduction

Person identification aims at verifying the individuals using their physiological or behavioral traits obtained from the data captured from various multimedia *e.g.*, color and/or depth information from single or multiple cameras [1]. Gait is one of the well-recognized biometric features which refers to identification of an individual using his/her walking style. Many existing biometric modalities, such as fingerprint, face, iris, earlobe geometry are being used in daily life and they require human interaction with the system [2, 3]. Conversely, gait can be acquired from a distance and also at low resolution in a non-invasive and hidden manner. Although gait may not be as powerful as the other biometric modalities are, its characteristics to recognize an individual from a distance and without any interaction with the system make it irreplaceable in many applications such as visual surveillance. However, gait recognition is often degraded by many covariates including clothing, shoes, walking surface, injuries, and viewpoints. Among these, the appearance changes in walk due to change in viewpoints are the most challenging and unavoidable in the practical surveillance systems. It introduces intra-personal variations which are always larger than inter-personal variations caused by other covariates [4].

In general, two different approaches are proposed for video-based gait recognition: appearance-based [4–15] and model-based [16–20]. The appearance-based approaches operate on the captured images directly. Usually, they extract human silhouettes from the video and generate a gait representation using the spatiotemporal shape and dynamic motion characteristics. Thus, the performance of such approaches depend on the accuracy of silhouette segmentation and hence an inaccurate segmentation (*e.g.*, due to clutter background) may lower the recognition accuracy [21, 22]. Though few researches *e.g.*, [23] proposed refinements in the segmented silhouette shape for gait recognition, however it is still a challenging problem in the literature. The model-based approaches uses the human body structure and motion models to identify the individuals. They characterize a human subject using a structural model and track several body-parts and joint positions over time to describe the gait using the underlying mathematical structure, *e.g.*, stick-figure [20], interlinked pendulum [19], and ellipse fitting techniques [16]; and the motion parameters of the subject, *e.g.*, joint angle trajectories [17], rotation patterns of hip and thigh [19], etc. However, the computation of such parameters require the localization of torso, which is difficult from the low-resolution images captured at distance in real surveillance systems. Although,

[*] Corresponding author
Email address: hassan.khan@uni-siegen.de (Muhammad Hassan Khan)

3D model-based approaches [20, 24] have demonstrated better cross-view recognition results due to their view-invariant nature but they require high-resolution images and they are computationally expensive [25]. This paper presents an appearance-based solution for cross-view gait recognition. However, it is worth mentioning that the proposed gait representation algorithm is different from the existing appearance-based approaches as it do not require any silhouette segmentation and the gait-cycle estimation.

Lately, the cross-view gait recognition has received significant research efforts due to its applications in surveillance systems. Numerous cross-view gait recognition techniques have been proposed which can be divided into three categories: (1) view-invariant gait descriptors [25–30], (2) construction of 3D gait descriptors [31–33], and (3) the view transformation-based features [4, 12, 13, 15, 34–38]. The first family of techniques develop a view-invariant gait descriptor by transforming the gait sequences of different views into a common space. These approaches perform well in specific scenarios and are hard to generalize for other cases. Moreover, their feature extraction phase is also disrupted due to self occlusion [4]. The approaches in the second category construct a 3D gait descriptor using multiple calibrated cameras. Such approaches perform good in a fully controlled and calibrated multi-camera environment, which is costly and computationally expensive [29]. The approaches in the third category construct a model to learn a mapping/projection of gait sequences perceived from multiple views. The cross-view descriptors are constructed using the learned model. In contrast to the first two categories, the approaches in the third category have demonstrated excellent recognition results. Moreover, they can be directly applied to the viewpoints which are significantly different from the side view, *e.g.*, frontal or back view. However, most of these techniques construct multiple mapping matrices, one for each pair of viewpoints.

The proposed method belongs to the third category. It constructs a deep neural network that learns a single model to transfer the knowledge of gait sequences from multiple viewpoints to one canonical view. Our learning scheme is based on the observation that the gait characteristics of a person from different viewpoints still exhibits a common structure which makes it different from others. Therefore, the gait related features should be separated from the viewpoint related features, which is not linearly possible. The proposed method works in three steps. First, a spatiotemporal gait representation is computed directly from the video sequences. Second, a deep neural network is trained which finds a shared high-level virtual path to map the gait descriptors from different viewpoints to a single canonical view. The spatiotemporal gait descriptor of side-view gait sequences are used as the canonical view. Third, the gallery and the probe sequences are transformed using the trained model in order to obtain their view-invariant gait representation and fed to subsequent classifier with their respective labels. We used a simple linear support vector machine (SVM) [39] as classifier. The major advantages of the proposed method are:

- Unlike the most existing cross-view gait recognition methods, *e.g.*, [4, 12, 29, 36, 40, 41] which require the silhouette segmentation to form a gait representation, the proposed method is based upon a spatiotemporal gait representation which is directly computed from the gait sequences. Thus, our method does not require the silhouette segmentation and the gait-cycle estimation.

- We learned a single model to map the gait sequences from all viewpoints to the canonical view using a pretty small set of multi-view video sequences of gait. Moreover, the proposed network does not require the information of viewpoints and other variations in the gait during the training of the network and at the construction of cross-view gait descriptors.

- The performance of the proposed algorithm is evaluated on two large benchmark cross-view gait databases: CASIA-B [42] and OU-ISIR large population (OULP) [43]. The recognition results and comparison with the state-of-the-art techniques confirm the effectiveness of the proposed method. It is worth mentioning here that our algorithm achieved excellent results using a simple linear support vector machine (SVM) [39] as classifier, which demonstrates the strength of proposed gait representation and the learned model based on deep neural network.

2. Related Work

2.1. Gait Representations

Numerous methods have been proposed to construct a gait representation from the images and video gait data. In general, they can be categorized into two groups: model-based approaches and appearance-based approaches.

Model-based methods aim to build a gait representation using the human body structure and motion models. Several human body parts and joint positions are tracked over time and used to identify the walkers. Lee et al. [16] proposed the modeling of human silhouette structure using seven different ellipses representing the various human body regions. They computed several statistical measurements on these regions over time to form a gait descriptor. The authors in [17] locate the joint locations and compute the joint angle trajectories at these locations to form a gait representation. Chai et al. [18] split the structure of human body region into three parts and the variance of these parts over time are combined to obtain a gait feature. In [19], a gait representation using the angular motion of the hip and thigh is presented. Recent studies [10, 20] have shown that these approaches highly depend on the localization of the torso, and require high-resolution images. They are also sensitive to video quality and are computationally expensive [10].

Appearance-based approaches do not build any structural or motion model, instead they operate on the recorded sequence of gait images directly. Usually, they extract human silhouettes from the images or video and drive various information for gait identification, *e.g.*, construct a template from silhouettes images [5, 6, 8], extract various gait parameters [7, 21], exploit shape [9] or projection analysis [44] on silhouettes. Most of them extract the human silhouettes from the images and combine them over the gait-cycle to obtain a template image which is used for person identification. Among them, gait energy image (GEI) [5] has been extensively used due to its simplicity and effectiveness. The representation is obtained by averaging the segmented silhouettes of a subject over the gait-cycle. The authors in [6] used the average of the difference images between two adjacent silhouettes as gait descriptor. Similarly, the authors in [8] used the average of silhouette body contours as gait representation. Goffredo et al. [7] proposed the use of height and width features from the normalized and scaled silhouette region to construct a gait representation. The authors in [21] used the radial basis function (RBF) network and deterministic learning on the height and width ratio and centroid of the contour to approximate an individual's gait. In [45], a comparative study of different convolutional neural network (CNN) architectures is proposed to recognize the gait. Different low-level features such as optical flow, gray pixels and depth maps are computed from the gait sequences; stacked them into spatio-temporal volumes separately and are passed to three different CNNs. Finally, their outputs are fused to obtain the gait signature. The authors in [46] proposed a framework for gait recognition using the combination of two CNNs which are modeled in one joint learning procedure and can be trained jointly. Besides, silhouette projection [44], shape analysis [9] and motion information [10, 47, 48] have been also exploited to construct a gait representation.

The appearance-based approaches are capable to recognize the individuals even from the low-resolution images and they are computationally efficient too compared to their counter part model-based approaches [10, 21, 49]. However, their recognition accuracy is highly based on the precise segmentation of silhouette from the background, which is still a challenging problem in literature. An inaccurate segmentation of silhouette shape may disrupt the construction of gait descriptor and degrade the recognition accuracy [21]. Conversely, the proposed method computes the spatiotemporal motion characteristics of an individual to represent his/her gait. It neither requires the silhouette segmentation nor the estimation of gait-cycle. Similar to our gait representation, the authors in [50, 51] computed motion descriptors from densely sampled points in a video sequence, and their higher level representation is obtained using histogram-based techniques. The aforementioned techniques have shown good recognition results when used within the same view scenario but performed rather poorly in cross-view situations because the gait recognition performance severely suffers from the appearance variance caused by the view change [4]. In the following section, we review the view-independent approaches to construct the cross-view gait descriptors.

2.2. Cross-view Gait Recognition

The existing cross-view gait recognition approaches can be categorized into three groups: view invariant gait features, 3D gait descriptors based approaches, and view transformation model.

The approaches in the first category construct a view invariant gait descriptor by transforming the gait sequences of different views into a common space. These approaches are geometry-based [26, 27], subspace learning-based [25, 28] and metric learning-based [28] approaches. The geometry based approaches construct a view-invariant gait descriptor using the geometrical properties of the gait sequences. For example, Kale et al. [26] proposed the perspective projection model to obtain a side-view gait images of an individual from any arbitrary viewpoint by assuming that the walking person is a 2D planar object in the sagittal plane. The authors in [27] proposed the transformation of motion trajectories from any arbitrary view to a standard plane and their similarities are compared to identify the individuals. The subspace learning-based approaches learn a joint subspace using the gait features from training data. The view-invariant features for testing sequences are obtained by projecting them on the learned subspace. The authors in [25] employ subspace learning and used direct linear discriminant analysis (DLDA) to create a single projection model for classification. Liu et al. [28] proposed to learn a joint subspace of the gait feature using joint principal component analysis (JPCA) to pair with different view angles. The authors in [15, 35] used the canonical correlation analysis (CCA) to project each pair of gait sequence into two subspaces with maximal correlation. However, these techniques construct multiple mapping matrices, *i.e.*, one for each pair of viewpoints. The authors in [29] proposed discriminative projection with list-wise constraints and rectification (DPLCR) to measure the similarity of gait features across the views.

The authors in [52, 53] proposed a method for activity-based person identification including walk using Fuzzy vector quantization. They exploited dynemes [54] to estimate the static body information, and the temporal information is preserved by calculating the similarity of each test body pose with all the dynemes. Later, a joint subspace is learned using linear discriminant analysis (LDA) to obtain a view-invariant representation of activities for classification. The method proposed in [53] employed the human body poses from different viewing angles to train a Self Organizing Map (SOM) network to determine the body pose prototypes which are subsequently used to describe the training actions by calculating the fuzzy similarities between all the prototypes and the human body poses appearing in each training action video. This action representation and their information is exploited to train two feed-forward neural networks for person identification and action recognition respectively. Metric learning-based approaches compute a weighting vector comprising the similarity score related to each feature, which is used to estimate the recognition score. The authors in [30] proposed the use of pairwise RankSVM algorithm [55] to improve the gait recognition with several variations, such as, view, clothing and carrying. The methods in [56, 57] learned the transformation matrices using a pair of gait features from two different viewpoints rather than learning a single mapping for all the viewpoints. Similarly, the technique in [58] proposed a tensor representation framework which employed coupled metric learning technique for cross-view gait recognition. They extract Gabor-based representations from GEIs of different views and project them to a common subspace for recognition. In [4, 59], a deep CNN using the GEI is proposed to measure the similarity between the gait features perceived from different viewpoints. Yan et al. [12] employed GEI with CNN to predict multiple attributes for cross-view gait recognition. These approaches perform well for limited scenarios, particularly when the view change is not large but usually they are hard to generalize for the other cases and their feature extraction phase is also disrupted due to self occlusion [4].

The approaches in the second category assume that the temporally synchronized images of a walking subject are available from multiple cameras. They constructs a 3D gait descriptor using multi-view synchronized images. Bodor et al. [31] proposed a 3D visual hull model to construct the view-invariant gait features using the input images from multiple cameras. In [32], a 3-D linear model is proposed to construct a view-independent gait feature using Bayesian rules. Zhao et al. [33] constructed a 3D human model for gait recognition using the video sequences captured from multiple viewpoints. The 3D gait descriptor-based techniques require costly setup of multiple calibrated cameras and they are computationally expensive. Moreover, they can only be used in a controlled environment and therefore they are considered unsuitable for real applications [25].

View transformation model (VTM) based approaches learn a mapping or a transformation relationship among the gait features perceived from different viewpoints. Later, the learned relationship is used to construct the cross-view gait descriptors prior to measure their similarity. These approaches can deal with view variations without relying on multiple cameras or camera calibration. Makihara *et al.* [38] proposed a singular value decomposition (SVD) based VTM to project frequency-domain based features from one view to another. Kusakunniran *et al.* [36] applied LDA for optimizing the gait features and to train a VTM for each pair of views. Instead of using SVD, [37] computes the local motion gait features and builds VTMs using support vector regression (SVR). Hu *et al.* [13] proposed a gait feature known as enhanced Gabor gait (EGG) which uses a non-linear mapping to encode the statistical and structural characteristics of gait across the views. They also exploit regularized local tensor discriminant analysis (RLTDA) to capture the nonlinear manifolds. In [40], a VTM using GEI and principal component analysis is proposed for view invariant feature extraction. Kusakunniran *et al.* [41] considered the VTM construction as a regression problem by adopting GEI and a Multi-Layer Perceptron (MLP) network to seek the motion information from source view which is used to estimate the gait in target view. The authors in [34] proposed a unitary linear projection to construct a cross-view gait descriptor. The technique proposed in [60] used a deep learning-based method to transform the gait descriptors of multiple viewpoints to a canonical view. They used a side-view GEI as canonical view and generative adversarial networks (GAN) as a regressor. In [61], a gait-related loss function is proposed for deep learning-based cross view recognition techniques to compute the discriminative features. It used spatial transformer network to localize the horizontal parts of walker, and long short-term memory model to encode their temporal attention. The method in [62] exploited deep learning model to compute the cross-view gait representation using a set of independent frames. It extracted local features from each silhouette and aggregated them into a single set-level feature which are mapped into a discriminative space to obtain the final representation.

In contrast to the previous two categories, VTM has shown the excellent recognition accuracy and can be directly applied to the viewpoints which are significantly different from the side view [60]. They are applicable to solve both cross-view and multi-view gait recognition problems and they do not require multiple calibrated cameras too. Moreover, they are computationally fast and therefore suitable for real-time applications [41]. However, the majority of existing methods in this category *e.g.*, [41, 60] use silhouettes of walkers to construct the gait features and build a VTM to transform the gait features from one view to another view. Other VTM based methods *e.g.*, [38, 40, 41] learned multiple transformations, one for each pair of viewpoints. That is, they can only transform from one specific viewpoint to another view. In contrast to the existing techniques, the proposed method do not require silhouette segmentation and trains a deep network to learn a single model for the transformation of knowledge from multiple viewpoints to a single canonical view and this model is used to construct the cross-view gait descriptors to identify the walker.

Preliminary results of this research are published in [63]. In this paper a number of improvements over [63] are proposed, including, a detailed literature review of the state-of-the-art gait recognition techniques and their classification into various categories, and a comprehensive explanation of the proposed architecture. More extensive experimental evaluations are performed. Two large benchmark cross-view gait databases are used to assess the performance of the proposed algorithm. Moreover, the recognition results are compared with several well-known gait recognition techniques.

3. The Proposed Technique

In the proposed gait recognition algorithm, a spatiotemporal gait representation [47] is computed from the video sequences. Then, a deep neural network is constructed to learn the transformation of spatiotemporal gait features from different source viewpoints to the canonical view. The learned model is used to construct the cross-view gait descriptors which are fed to the subsequent classifier.

3.1. Spatiotemporal based Gait Representation

The proposed method constructs a cross-view gait descriptor using the spatiotemporal features of gait to characterize the distinct motion information of individuals. We recall that these features are extracted directly from the video sequences within a space-time volume [64] and do not involve any human-body segmentation from the background. We computed several local motion descriptors such as Motion Boundary Histogram (MBH) [65] and Histogram of Oriented Gradient (HOG) [66] from the video sequences to construct a gait representation. The HOG represents the person's static appearance by capturing the magnitude and gradient information from the gait images. Similarly, the MBH is constructed by computing the gradients over the horizontal and the vertical components of optical flow and encodes the relative motion information between the pixels in the direction of respective axis. The evaluation of the various motion descriptors demonstrated that HOG and MBH together outperform the others features for gait recognition [67]. Since, the HOG comprises the person's static appearance and the MBH highlights the information about the changes in the optical flow field (*i.e.*, motion boundaries), therefore when used collectively they have a greater impact in identifying a person using his/her appearance and local motion characteristics. Therefore, we selected HOG and MBH as local descriptors for gait recognition.

Usually, the local descriptors are used to build a signature (*i.e.*, feature encoding) to characterize an image or video sequence. In this work, the local motion descriptors are encoded using Fisher vector [68] encoding and a codebook based on Gaussian mixture model (GMM). The GMM describes the distribution over feature space and can be expressed as,

$$p(\mathscr{X} \mid \theta) = \textstyle\sum_{k=1}^{K} w_k \mathscr{N}(x \mid \mu_k, \sum_k), \tag{1}$$

where K are the number of components (*i.e.*, clusters) in GMM and $\theta = \{w_k, \mu_k, \sum_k | k = 1, 2, \ldots, K\}$ represents the set of model parameters with w_k is the weight, μ_k is the mean vector and $\sum_k$ is the covariance matrix of the $k-$th cluster. Moreover, $\mathscr{N}(x \mid \mu_k, \sum_k)$ is the D-dimensional Gaussian distribution. Each mixture is assumed to represent a specific appearance and motion pattern shared by the local descriptors. For a given feature set $\mathscr{X} = \{x_t | t = 1, \cdots, \mathscr{T}\}$, the soft assignment of data x_t to cluster k is defined as,

$$q_t(k) = \frac{w_k \mathscr{N}(x_t \mid \mu_k, \sum_k)}{\sum_{j=1}^{K} w_j \mathscr{N}(x_t \mid \mu_j, \sum_j)} \tag{2}$$

The Eq. (2) assigns the local descriptor to multiple clusters in a weighted manner using the posterior component probability given by the descriptors. We used randomly selected one million local descriptors from the training set to build a codebook with GMM. The number of components K in GMM are empirically computed and set to 2^8. The Fisher vector (FV) representation consists of the deviation of the local descriptor from the generative model. This deviation can be computed using the gradient of the descriptor log-likelihood with respect to the model parameters θ [68]. That is, the gradient vector with respect to mean μ_k and covariance $\sum_k$ is computed as:

$$u_k = \frac{1}{\mathscr{T}\sqrt{w_k}} \sum_{t=1}^{\mathscr{T}} q_t(k) \frac{x_t - \mu_k}{\sum_k} \tag{3}$$

$$v_k = \frac{1}{\mathscr{T}\sqrt{2w_k}} \sum_{t=1}^{\mathscr{T}} q_t(k) \left[\frac{(x_t - \mu_k)^2}{\sum_k^2} - 1 \right], \tag{4}$$

The Fisher encoding for the set of local descriptors $\mathscr{X}$ is computed by concatenating the u_k and v_k for all $k = 1, 2, ..., K$ components. That is,

$$f = \left[u_1^\top, v_1^\top, u_2^\top, v_2^\top, \ldots, u_K^\top, v_K^\top\right]^\top$$

The MBH$_x$, MBH$_y$ and HOG descriptors are encoded as described above and they are fused using the representation level fusion [69]. The length of each spatiotemporal gait feature is set to $2,000$ using the principal component analysis (PCA). The proposed gait features demonstrated excellent performance in the lateral view gait recognition [47], which encouraged us to choose this gait representation to construct the cross-view gait descriptor.

3.2. View Transfer Model

A non-linear deep neural network (NDNN) is proposed to transfer the gait descriptors of different viewpoints to a single canonical view. Most of the existing cross-view gait recognition approaches *e.g.*, [13, 15, 32, 34, 36, 38] use a set of linear transformations to map the source gait descriptors to the target viewpoint. Therefore, such approaches are unable to capture the non-linear manifolds where the realistic gait scenario may lie [4]. To overcome this problem, the proposed NDNN learns a multi-step virtual path between the source viewpoints and their respective canonical view. The source viewpoints are mapped to some intermediate virtual views along the non-linear path prior to construct the final canonical view. Our learning scheme is based on two observations: first, the gait representation of a person from different viewpoints has the same high-level characteristics that make it different from the others. Second, there would be m different virtual paths connecting

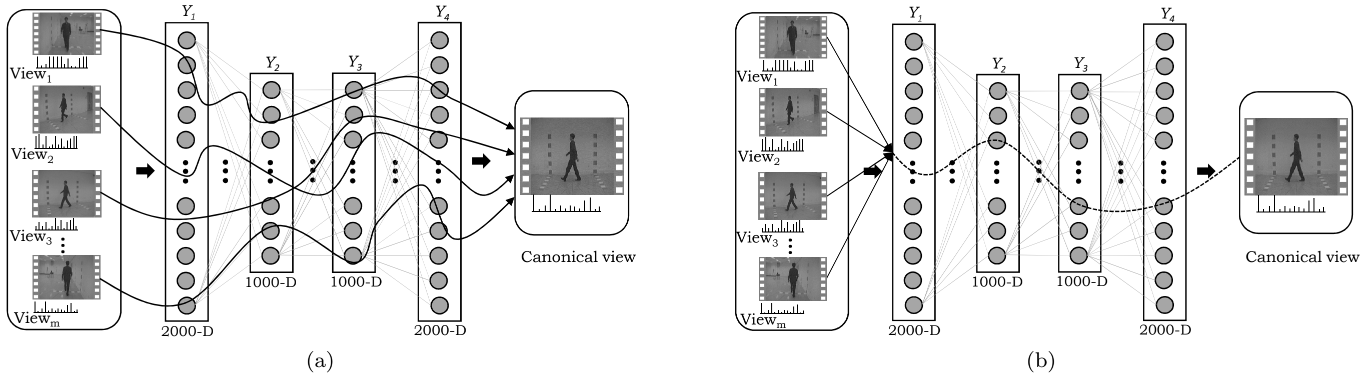


Figure 1: Training of the proposed non-linear deep neural network (NDNN). (a) The gait video sequences observed from unknown viewpoints are transformed to the canonical view. (b) The proposed network forces the m different virtual paths to learn a single, high-level and shared virtual path (dotted line) which connects all the source viewpoints to the canonical view.

m different source viewpoints to the respective canonical view [70], and the proposed network aims to force them to learn a single non-linear virtual path. That is, the proposed network starts with m different virtual viewpoints and force them to agree on a single virtual path as shown in Fig. 1.

The proposed NDNN consists of 4 layers with H hidden units in each layer. Let a layer L with H hidden units be represented as L_H. During training, the output of each layer is forwarded as an input to the next consecutive layer. In a given training instance $X_{ij} \in \Re$, where X_{ij} is the j-th training instance from i-th viewpoint, the output of the first layer is computed as,

$$Y_1 = f(W_1 X + b_1), \tag{5}$$

where Y_1 represents the output of first layer, W_1 is the weight matrix and b_1 is the bias vector to be learned for the first layer. Moreover, $f(\cdot)$ represents the non-linear activation function which usually includes hyperbolic tangent function $\tanh(x)$, the logistic function $f(x) = \frac{1}{(1+e^{-x})}$, and the rectified linear unit (ReLU) $f(x) = \max(0, x)$. However, we used the leakyReLU (Leaky Rectified Linear Unit) as an activation function. It is a variant of ReLU, perform efficiently and unlike other activation functions such as sigmoid, it does not suffer from vanishing gradient problem. In comparison with ReLU, it allows a small non-zero gradient when the unit is saturated and not active [71]. Specifically, it assigns a small slope to the negative part instead of completely dropping it. That is,

$$f(x) = \begin{cases} x & \text{if } x > 0; \\ \alpha x & \text{otherwise} \end{cases} \tag{6}$$

where α is a small constant. The output of first layer Y_1 is forwarded as an input to the second layer, which would be processed as,

$$Y_2 = f(W_2 Y_1 + b_2), \tag{7}$$

where Y_2 denotes the output of the second layer, W_2 is the weight matrix, and b_2 is the bias vector to be learned for the second layer. The output of the last fully connected layer is computed as

$$t(X_{ij}) = Y_L = f(W_L Y_{L-1} + b_L), \tag{8}$$

where $t(X_{ij})$ represents the non-linear transformation of X_{ij}, determined by the parameters W_L and b_L. The input to our proposed NDNN are the spatiotemporal gait descriptors (Section 3.1) from the source viewpoints and the output is the spatiotemporal gait descriptors from the canonical view. We recall that the spatiotemporal gait descriptors of side-view gait sequences are used as canonical view. Therefore, the proposed network tries to get the output of $t(X_{ij})$ to be close to its canonical view (*i.e.*, X_{cj}) regardless of its input view as described earlier.

We empirically chose the appropriate number of units in each layer to ensure that the network can efficiently learn the underlying structure of the data. The sizes of first and last layers are set to $2,000$ due to the dimension of the spatiotemporal gait descriptors and the size of two intermediate layers is set to $1,000$ for each. Since the input dimension is $2,000$, the redundant information in the input features is removed by mapping it to a high-level but low-dimensional representation, and this computation is performed in the first two fully connected layers. Later, this low-dimensional compact representation is mapped back to a high-dimensional output layer using the last two fully connected layers as shown in Fig. 1. The output of the last layer is the canonical representation of the input features.

The learning of the proposed network consists of minimizing the loss function of the reconstruction error over all training samples by updating the set of parameters $\theta = \{W_L, b_L | L = 1, 2, \ldots, 4\}$. Moreover, we added weight decay J_w regularization to penalize the objective function in order to reduce the effect of over-fitting. For m viewpoints and n instances in each view, the reconstruction

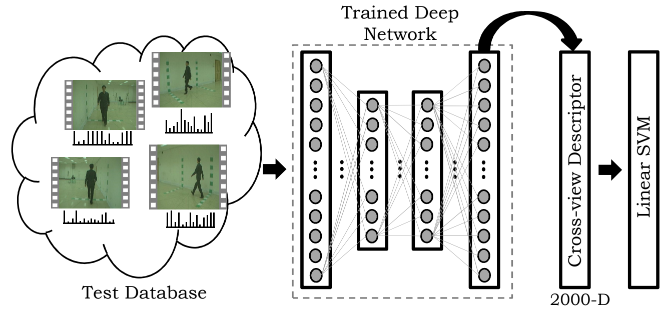


Figure 2: Construction of cross-view descriptor and classification using Linear SVM. The output of last layer is used as cross-view gait descriptor.

error e_Θ is:

$$e_\Theta = \frac{1}{2mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \parallel X_{cj} - t(X_{ij}) \parallel^2 + \lambda_w J_w, \qquad (9)$$

where $i = 1, \ldots, m$ are the viewpoints, $j = 1, \ldots, n$ are the training instances, c is the canonical view, λ_w is represents the weight decay parameter and $J_w = \sum_{h=1}^{H} W_h^2$ is L$_2$ regularization. Weight decay penalizes large weights more strongly which may cause highly curved and non-smooth mappings [72]. The weight decay penalizes the large weights and keeps them small to make the mappings smooth and to reduce the over-fitting. The performance of the proposed NDNN is also analyzed using Huber loss (also known as smooth-L$_1$) function [73] in the training process. It computes the element-wise difference and uses its squared term if the error falls below a threshold which is set to 1, and absolute L$_1$ distance otherwise. It is less sensitive to outliers as compare to mean square error which simply square the difference and may result in exploding gradients. However, we did not observe any improvement in the training accuracy and this could be due to the nature of local descriptors which are used in gait encoding. These descriptors are normalized with L$_2$-norm to make sure that they take similar ranges of values and are not too large. We used mini-batch stochastic gradient descent method through back-propagation to minimize the objective function over all training samples.

3.3. Cross-view gait descriptor and classification

We consider the $t(X_{ij})$ (Eq. (8)) as the non-linear transformation function which transforms X_{ij} to its respective canonical view X_{cj}. In particular, this function provides the canonical view representation of a gait sequence obtained from any unknown viewpoint. It is worth mentioning that the proposed network does not require viewpoint information during the training of the network and at the construction of cross-view gait descriptors. Therefore, the gait sequences from the testing dataset are propagated through the trained network and the final output of the network, which consists of a set of non-linear transformations $Y_1, Y_2, \ldots, Y_4$ from source viewpoint to the canonical viewpoint (Figs. 1 and 2), is chosen as the cross-view gait descriptor because it encodes the influence of all these transformations and provides the canonical view representation of gait regardless of its input view. Let $X'_{ij} \in \Re$ be the j-th testing instance from any unknown i-th viewpoint, the final cross-view gait descriptor is constructed as,

$$t(X'_{ij}) = f(W_L \ldots f(W_2(f(W_1 X'_{ij} + b_1)) + b_2) + \ldots + b_L), \qquad (10)$$

where $t(X'_{ij}) \approx X'_{cj}$. For classification, the gait sequences of the gallery set are selected and their cross-view gait descriptors are obtained using the learned NDNN (*i.e.*, by propagating them through the trained network) which are used to train a classifier along with their respective labels. At testing, the cross-view descriptors of the probe sequences are computed using Eq. (10) and fed to SVM to identify the walker. We used a simple linear SVM [39] for this purpose and it achieved excellent results which shows the strength and robustness of our deep network. The distribution of the gallery (to train a SVM) and the probe sets is outlined in the experimental evaluation of the respective gait database.

Table 1: Proposed gait descriptor dimension selection: recognition results (%) on CASIA-B gait database using different values of PCA in feature encoding.

Dimension	1,500	2,000	4,000	8,000	16,000
Accuracy	75.15	83.20	75.90	64.70	44.80

4. Experiments and Results

The proposed gait recognition algorithm is evaluated on two large benchmark cross-view gait datasets: CASIA-B [42] and OULP [43]. Several experiments are performed and the recognition results are compared with the existing state-of-the-art methods.

4.1. Implementation details

The first step in the proposed algorithm is to compute the gait descriptors. The spatiotemporal gait representation is computed from all the walking video sequences using the algorithm described in Section 3.1. The proposed NDNN is implmented using Keras framework [74] with Tensorflow [75] as back-end, and trained using the RMSprop variation of the gradient descent algorithm [76] with its default parameters (*e.g.*, learning rate is initialized with 0.001), for 150 epochs. The weights are updated with the mini-batch of size 33. We used LeakyReLU as the activation function with $\alpha = 0.01$, and Mean Squared Error function as the loss function for the model. The network is trained using back-propagation with logistic regression loss. Due to small number of layers, the network parameter $\theta = \{W_L, b_L \mid L = 1, 2, ..., 4\}$ is initialized using simple random initialization method [72]. In particular, all the bias terms b_L are initialized with zero and weight matrix W_L is initialized using a Gaussian distribution with zero mean and 0.05 standard deviation. The values of the λ_w is set to 0.0001. We used multi-resolution search [72] to find the optimal values of the hyper-parameters of NDNN. That is, first the parameter values are tested from a larger range and few best configurations are selected. Then, a narrow search space is exploited around these values to select the optimal values in the second step. For example, as suggested by [77], we evaluated the performance of the network with increasing number of hidden layers and stopped when obtained a peak performance on validation data. It is empirically concluded that increasing the further number of hidden layers beyond two did not improve the performance. Similarly, the hidden layer sizes are evaluated in the range $[500, 1,500]$. We train the network in a regular way using a gradient descent based algorithm, and then the gallery and probe gait sequences are propagated through the trained network in order to obtain their cross-view gait representation (*i.e.*, an approximation of the respective canonical view). Our cross-view gait representation is a $2,000$ dimensional descriptor, the value is selected empirically. We evaluated the performance of the proposed method on CASIA-B dataset using different values of PCA. During classification, the gallery set (θ_g) contains the gait sequences of viewpoint 90° and the rest of the viewpoints are used in probe sets (θ_p) separately. The average recognition accuracies across all the viewpoints are presented in Table 1. The experimental evaluation reveals that the best results are achieved when the descriptor dimension is $2,000$.

The computational complexity of the proposed algorithm is computed using the gait sequences of CASIA-B gait database. We carefully analyzed the computation time of each step involved in the proposed algorithm. The local descriptors are extracted from the entire video sequence on full resolution (*i.e.*, 320×240) using 50 frames. This step is computationally the most expensive and requires on average 2.4 seconds. The reported time is highly dependent on the frame's resolution and the length of gait sequence which can be optimized considering these factors. The feature encoding and its cross-view gait representation is obtained in 0.2 seconds, and the final classification step took 0.02 seconds. The experimental evaluation is carried out on a machine with an Intel i7-6700K CPU, 64GB RAM and a NVidia GTX TITAN X GPU.

4.2. Performance evaluation on CASIA-B dataset

The CASIA-B database comprises the gait video sequences of 124 subjects. The gait sequences are collected in a controlled indoor environment using eleven different viewpoints: $0°, 18°, 36°, \ldots, 180°$. Sample images from each viewpoint are shown in Fig. 3 to demonstrate the appearance-changes due to change in the viewing angle. The database contains ten walk sequences for each subject with three variations, namely: normal walk (*nm*), walk with bag (*bg*), and walk with coat (*cl*). Among these ten sequences, six belong to *nm* and two to each *bg* and *cl*. Fig. 4 presents sample images from viewing angle 90° to demonstrate the gait under various conditions. The viewpoint 90° is selected as a canonical view during the network training.

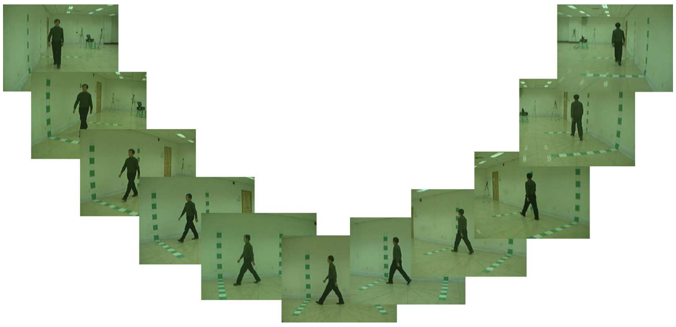
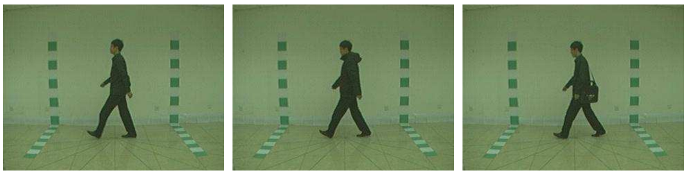


Figure 3: Sample images of CASIA-B dataset from each viewpoint 0° to 180° (left to right).



(a) (b) (c)

Figure 4: Sample images from CASIA-B dataset demonstrating the variations in walk. (a) Normal walk, (b) walk with coat, and (c) walk with bag.

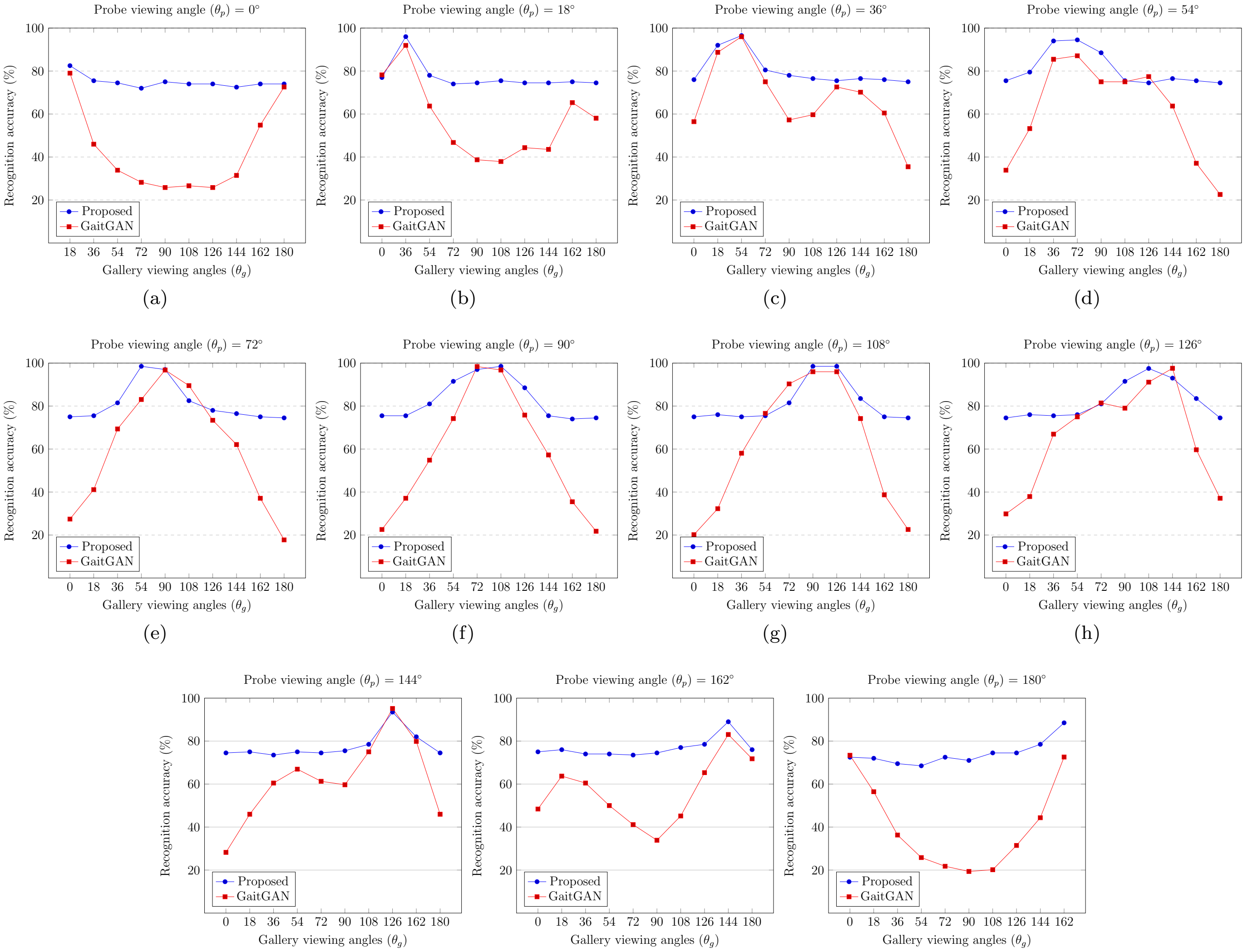
Cross-view gait recognition on CASIA-B database is challenging, particularly when the cross-view angle is large. It is even more difficult when the probe and gallery sets belong to different walking scenarios [25]. Similar to [4, 15], the normal gait sequences of 24 subjects are randomly chosen to train the proposed deep network and the remaining gait sequences of 100 subjects are used to evaluate the performance of the cross-view gait recognition algorithms. In all the experiments, the first four normal walk sequences (*i.e.*, $nm_1 - nm_4$) of 100 subjects in the dataset are used to form a gallery and the rest are used in different probe sets. Similar to the recent state-of-the-art techniques [4, 25, 60], three different types of experiments are performed and the achieved results are compared with the existing techniques.

In the first set of experiments, we evaluated the cross-view gait recognition performance of the proposed NDNN on each view-pair. That is, for each experiment the gallery and the probe gait sequences are selected from different viewpoints. We first construct a gallery set (θ_g) by taking each view iteratively from $\{0°, 18°, \ldots, 180°\}$ and construct a probe set (θ_p) from the rest of the ten views, separately, excluding the identical view in the gallery set. Fig. 5 presents the performance of proposed NDNN in each experiment. The plots show that our method achieves better performance than similar GaitGAN [60] algorithm.

Most existing approaches [4, 15, 25, 29] report perfor-

Table 2: Comparison of recognition results (%) on CASIA-B gait database with gallery view θ_g is 90° during SVM classification. The best results are marked in bold.

$\theta_p : nm_5 - nm_6$	40°	18°	36°	54°	72°	108°	126°	144°	162°	180°
JDLDA [25]	20	25	37	58	94	-	-	-	-	-
MvDA [78]	17	27	36	64	95	-	-	-	-	-
JSL [28]	20.5	35.5	56.5	81.5	96.5	96	89.5	50	34.5	21.5
Method [15]	-	-	42	70	95	96	70	41	-	-
Method [42]	0.4	2.4	4.8	17.7	82.3	82.3	15.3	5.2	3.6	1.2
GEI+TSVD [36]	15	18	22	52	-	90	55	-	20	10
GaitGAN [60]	22.58	37.1	54.8	74.2	**98.4**	96.8	75.8	57.3	35.5	21.8
Ours	**75**	**74.5**	**78**	**88.5**	97	**98.5**	**91.5**	**75.5**	**74.5**	**75.5**



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k)

Figure 5: The recognition accuracy (%) of the proposed NDNN and GaitGAN [60] using the gallery set (θ_g) from all the 11 views, separately. The probe viewing angles are (a) 0°, (b) 18°, (c) 36°, (d) 54°, (e) 72°, (f) 90°, (g) 108°, (h) 126°, (i) 144°, (j) 162°, and (k) 180°.

Table 3: Comparison of recognition results (%) on CASIA-B gait database. The results are obtained by averaging the accuracies on all gallery views except the identical view in the probe set. The best results are marked in bold.

$\theta_g : nm_1 - nm_4$	$0° - 180°$				$36° - 144°$		
$\theta_p : nm_5 - nm_6$	0°	54°	90°	126°	54°	90°	126°
SVR [37]	-	28	29	34	35	44	45
TSVD [36]	-	39	33	42	49	50	54
Method [15]	46.3	52.4	48.3	56.9	65	67.8	69.7
ViDP [34]	-	59.1	50.2	57.5	83.5	76.7	80.7
EGG-RLTDA [13]	-	-	-	-	69.8	74.4	73.9
JDLDA [25]	-	27.16	25.7	29.9	39	40.9	44.6
GII [29]	-	63	55	62.1	80.1	78	80
GaitGAN [60]	41.9	64.5	58.1	65.7	78.6	77.3	81.7
CNN [4]	54.8	77.8	64.9	76.1	**90.8**	85.8	**90.4**
Ours	**74.8**	**80.9**	**83.2**	**84**	83.9	**88.7**	85.8

mance on 54°, 90° and 126° views in a probe set, separately, we selected the same views for performance comparison. The recognition results are presented in Table 3, where the gallery ($0° - 180°$) represents the average performance for all gallery viewing angles ranging from 0° to 180° except the identical viewpoint which is used in the probe set. Similarly, the gallery ($36° - 144°$) represents the average performance for all gallery viewpoints ranging from 36° to 144° except the identical viewpoint which is used in the probe set. The results show that the proposed method outperforms the compared methods in most experiments.

In the second set of experiments during SVM classification, according to the recommendations of [15, 28] the gallery set (θ_g) contains the gait sequences of viewpoint 90° and the rest of the viewpoints are used in probe sets (θ_p) separately. We recall that the gait sequences of all different viewpoints are used to train the proposed NDNN while the viewpoint 90° is selected as canonical view during the network training. The recognition results and comparison with the existing methods are presented in Table 2. The results show that the proposed method outperforms the state-of-the-art in all experiments except at viewing angle 72° where GaitGAN [60] performs slightly better than our method. It may be noted that in some cases the proposed NDNN performs exceptionally well compared to the existing methods. For example, it achieves superior performance over the state-of-the-art when the gallery and probe viewpoints are significantly different from each other (*e.g.*, $\theta_g = 90$ and $\theta_p = 0$) see Table 2.

The GaitGAN [60] method exploits GEI as gait representation and learns a VTM using GAN to learn a single model for mapping of different viewpoints. However, their performance degrades heavily when the gallery and probe viewing angles are significantly different from each other, as shown in Fig. 5 and Table 2. The superior performance of the proposed NDNN comes from two factors, an effective gait representation to encode the within-class geometry and the efficient view transformation model, while the competing methods (*e.g.*, [15, 36]) incorporate the information from limited viewing angles.

Table 4: Comparison of recognition results (%) on CASIA-B gait database under various conditions. In each block, two recognition results of *bg* and *cl* are presented with the separation of '/'. The best results are marked in bold.

θ_p	54° (*bg*/*cl*)	90° (*bg*/*cl*)	126° (*bg*/*cl*)
Method [79]	76.4 / **87.9**	73.7 / **91.1**	76.9 / **86.2**
RLTDA [13]	80.8 / 69.4	76.5 / 72.1	72.3 / 64.4
R-VTM [40]	40.7 / 35.4	58.2 / 50.3	59.4 / 61.3
FT-SVD [38]	26.5 / 19.8	33.1 / 20.6	38.6 / 32.0
CNN [4]	92.7 / 49.7	88.9 / 75.6	86.0 / 51.4
Method [42]	24.2 / 16.5	44.0 /27.8	31.9 /18.1
GaitGAN [60]	53.2 / 43.6	62.1 / 43.6	66.1 / 41.1
Ours	**93** / 84.5	**94.5** / 88.5	**91.5** / 82.5

In the third set of experiments, the robustness of the proposed method is evaluated under various conditions on the gait sequences recorded at 54°, 90° and 126°. Similar to [4, 79], the gallery set comprises the gait sequences of viewing angle 36°, 108° and 144° and the probe consists of 54°, 90° and 126°, respectively. The results are documented in Table 4. The statistics show that our method outperforms the compared methods in most experiments.

Next we performed a set of experiments using different number of instances in network training as carried out in [4]. The similar configuration of the network parameters is selected as reported in first set of experiments except the number of training instances. In particular, the normal walk sequences of 74 subjects are used to train the proposed NDNN and the rest normal gait sequences of 50 subjects are used to evaluate the performance of the cross-view gait descriptors. We named this experiment as similar-walk-condition [4]. The recognition accuracy of our method increases by 5% when 74 subjects are used for training, while the method in [4] reported an increase of 20% in recognition accuracy. The comparison of recognition results using 24 and 74 subjects in the network training is presented in Figs. 6a and 6b, respectively. In case of cross-walk-condition as outlined in third set of experiments, the gallery set comprises the normal gait sequences of viewing angle 36°, 108°and 144°whereas, the probe set consists of walking sequences of person carrying a bag during the walk (*bg*), and the sequences of walker wearing a coat (*cl*)under viewing angle 54°, 90°and 126°, separately. The recognition results under both conditions (*i.e.*, *bg* and *cl*) are presented in Figs. 6c and 6d, respectively. The proposed method is proven to be robust under cross-walk-condition experiments and achieves better recognition accuracy than method in [4], the difference is up to 20%. The results of these experiments reveal that the our method outperforms in most of the experiments in both similar and cross walking conditions, even when the network is trained using limited number of subjects *i.e.*, 24. Achieving high detection accuracy under limited training dataset is no doubt a challenging task and the performance of the proposed method in this scenario is appreciable. Furthermore, another advantage of the proposed method is its

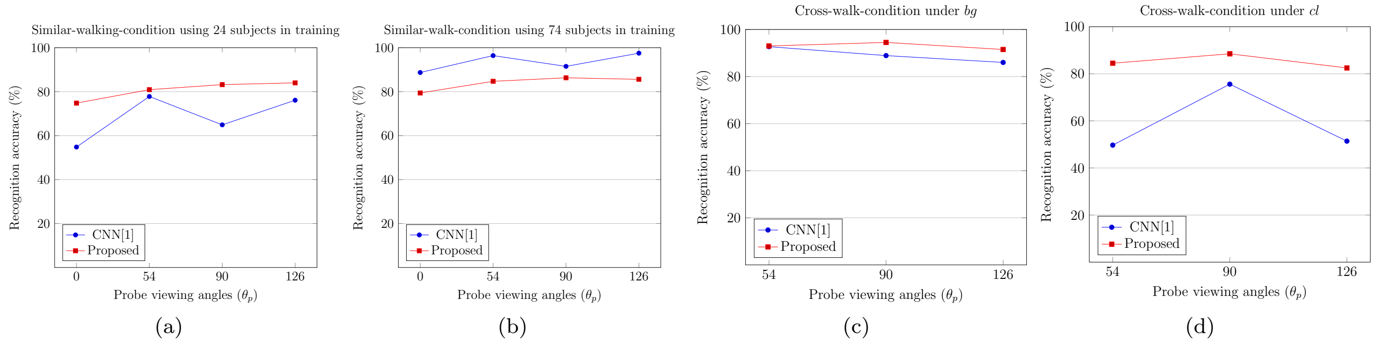


Figure 6: The recognition accuracy (%) of the proposed NDNN and CNN [4]. (a-b) The results are obtained under similar-walking-condition by averaging the accuracies on all gallery views except the identical view in the probe set. (c-d) The results are obtained under cross-walking-condition.

simple gait representation which neither requires the silhouette segmentation nor gait-cycle estimation, and it can be computed directly from the gait video sequences.

4.3. Performance evaluation on OULP dataset

The OU-ISIR large population (OULP) is the largest cross-view gait database which comprises the gait sequences of more than $4,000$ subjects. The gait sequences are recorded in an indoor environment at 30 frames-per-second, under four viewing angles: $55°, 65°, 75°$ and $85°$. Each subject was asked to walk along a course twice in a natural manner. Fig. 7 presents some example images of a walking subject with four viewing angles. The viewpoint $85°$ is selected as a canonical view during the network training. The researches *e.g.*, [59, 80, 81] designed two different types of experiments on this dataset, same-view experiment and cross-view experiment. In the same-view experimental setting, the gallery and probe gait sequences belong to the same viewing-angle, while in the cross-view setting they belong to different viewing-angles. Particularly, we first construct a gallery by picking each view iteratively from $\{55°, 65°, 75°, 85°\}$ and construct a probe from the rest of the three views separately.

The OULP dataset is divided into 5 parts and the distribution is publicly available at[1]. Each part contains the division of $1,912$ subjects in two equal disjoint sets. A five 2-fold cross validations are performed in both type of experiments to meet the protocols of benchmarks as in [59, 80–82]. That is, for each part when the gallery and the probe sets are exchanged 2-fold cross validation is adopted. In each run, we consider one set as testing, and train a network with the remaining set and vice versa. The average recognition accuracies obtained by the proposed method and other compared methods are reported in Table 5. The recognition results reveal that our method outperforms the compared methods in most experiments.

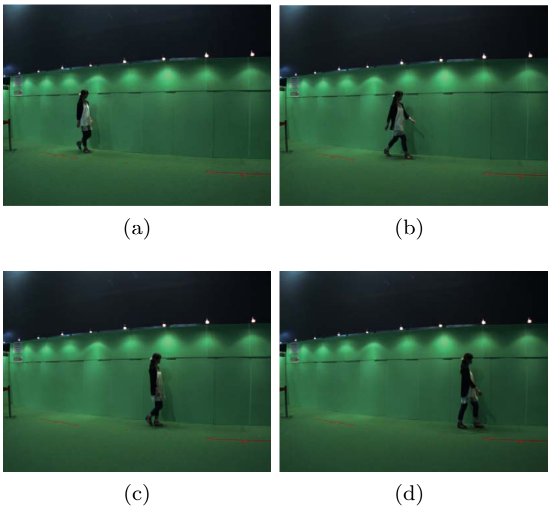


Figure 7: Sample images from OULP gait dataset. (a) $55°$, (b) $65°$, (c) $75°$, and (d) $85°$.

One can conclude from the results presented in Tables 3 and 5 that the proposed method performed consistently better than state-of-the-art cross-view gait recognition techniques in many experiments. We observed that the superior performance of the proposed algorithm is due to two factors, the effective gait representation and the efficient view transformation model. The proposed gait representation is based upon the spatiotemporal characteristics of gait, in particular, it comprises of two features, one captures the static appearance of the individual and the other considers the local motion information extracted from the changes in the optical flow field. To support this argument, an investigation study is also carried out to evaluate the performance of the proposed algorithm using the gait representations proposed in [50] and [51]. The gait features computed using [50, 51] are used as input to proposed NDNN framework. The experimental evaluation showed that the recognition accuracy of the proposed algorithm is drastically degraded up to 26% when the gait descriptors of [50, 51] were used as input to network. Though,

[1] http://www.am.sanken.osaka-u.ac.jp/BiometricDB/dataset/GaitLP/Benchmarks.html

Table 5: Performance evaluation (%) and comparison with the existing methods on OULP gait dataset under same and cross view settings. The values in parenthesis () represent the recognition results within the same-view settings. The best results are marked in bold. Note that hyphen (-) means that either the results are not available or the respective approach cannot be evaluated under those experiments.

Gallery	Method	probe			
		55°	65°	75°	85°
55°	DeepGait [80]	(97.4)	96.1	93.4	88.7
	wQVTM [82]	-	78.3	64.0	48.6
	GEINet [59]	(94.7)	93.2	89.7	79.9
	Method [81]	(95.2)	93.6	81.2	62.2
	PdVS [83]	-	76.2	61.4	45.5
	AVTM [83]	-	77.7	64.5	42.7
	Method [4]	(98.8)	**98.3**	**96.0**	80.5
	Ours	**(100)**	95.1	94.9	**97.5**
65°	DeepGait [80]	**97.3**	(97.6)	97.2	95.4
	wQVTM [82]	81.5	-	79.2	67.5
	GEINet [59]	93.7	(95.1)	93.8	90.6
	Method [81]	90.9	(95.3)	95.5	90.2
	PdVS [83]	76	-	77.1	65.5
	AVTM [83]	75.6	-	76.4	62.8
	Method [4]	96.3	(98.9)	**97.3**	83.3
	Ours	94.8	**(100)**	95.5	**97.5**
75°	DeepGait [80]	93.3	97.5	(97.7)	97.6
	wQVTM [82]	70.2	80.0	-	78.2
	GEINet [59]	90.1	94.1	(95.2)	93.8
	Method [81]	77.5	94.4	(96.0)	96.0
	PdVS [83]	60.3	76.2	-	76.5
	AVTM [83]	59.9	74.9	-	76.3
	Method [4]	94.2	**97.8**	(98.9)	85.1
	Ours	**95.1**	96.0	**(100)**	**97.7**
85°	DeepGait [80]	89.3	96.4	98.3	(98.3)
	wQVTM [82]	51.1	68.5	79.0	-
	GEINet [59]	81.4	91.2	94.6	(94.7)
	Method [81]	55.4	87.1	94.8	(94.7)
	PdVS [83]	40.5	60.6	73.1	-
	AVTM [83]	40.2	61.9	74.3	-
	Method [4]	90.0	96.0	**98.4**	(98.9)
	Ours	**98.0**	**97.1**	97.7	**(100)**

these feature are also computed using local spatiotemporal descriptors within a space-time volume along the trajectories, however several features are exploited to construct the final gait representation. Moreover, the view transformation model is built through a non-linear deep neural network which is able to capture the non-linear manifolds in gait and learns multi-step virtual path between the source viewpoints and their respective canonical view. Due to these advantages, our method achieved higher cross-view gait recognition accuracies than the existing methods.

5. Conclusion

In this paper, a novel cross-view gait recognition algorithm is proposed using a non-linear deep neural network. Spatiotemporal motion cues of human walk are extracted and used to construct a gait descriptor. The network is trained using these gait descriptors, which transfers the knowledge of an individual's gait sequences from different viewpoints to a single canonical view. The cross-view gait representation of a testing instance is achieved by a forward propagation through the trained network. A simple linear support vector machine is used to classify the cross-view descriptors. The experimental evaluations performed on two benchmark cross-view gait datasets and comparisons with the state-of-the-art methods confirm the effectiveness of the proposed algorithm.

References

[1] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, R. Hu, Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing, IEEE Trans. Multimedia 18 (12) (2016) 2553–2566.

[2] Y. Sun, M. Zhang, Z. Sun, T. Tan, Demographic analysis from biometric data: Achievements, challenges, and new frontiers, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2018) 332–351.

[3] J. K. Pillai, M. Puertas, R. Chellappa, Cross-sensor iris recognition through kernel learning, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 73–85.

[4] Z. Wu, et al., A comprehensive study on cross-view gait based human identification with deep CNNs, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 209–226.

[5] J. Man, B. Bhanu, Individual recognition using gait energy image, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 316–322.

[6] E. Zhang, Y. Zhao, W. Xiong, Active energy image plus 2dlpp for gait recognition, Signal Processing 90 (7) (2010) 2295–2302.

[7] M. Goffredo, J. N. Carter, M. S. Nixon, Front-view gait recognition, in: Int. Conf. Biometrics: Theory, Applications and Systems, IEEE, 2008, pp. 1–6.

[8] C. Wang, et al., Human identification using temporal information preserving gait template, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2164–2176.

[9] W. Kusakunniran, et al., Pairwise shape configuration-based psa for gait recognition under small viewing angle change, in: Int. Conf. Advanced Video and Signal-Based Surveillance (AVSS), IEEE, 2011, pp. 17–22.

[10] Y. Yang, D. Tu, G. Li, Gait recognition using flow histogram energy image, in: Proc. Int. Conf. Pattern Recognit. (ICPR), 2014, pp. 444–449.

[11] S. Yu, H. Chen, Q. Wang, L. Shen, Y. Huang, Invariant feature extraction for gait recognition using only one uniform model, Neurocomputing 239 (2017) 81 – 93.

[12] C. Yan, B. Zhang, F. Coenen, Multi-attributes gait identification by convolutional neural networks, in: Int. Congress on Image and Signal Processing (CISP), IEEE, 2015, pp. 642–647.

[13] H. Hu, Enhanced gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition, IEEE Trans. Circuits Syst. Video Technol. 23 (7) (2013) 1274–1286.

[14] W. Xu, C. Luo, A. Ji, C. Zhu, Coupled locality preserving projections for cross-view gait recognition, Neurocomputing 224 (2017) 37 – 44.

[15] W. Kusakunniran, et al., Recognizing gaits across views through correlated motion co-clustering, IEEE Trans. Image Process. 23 (2) (2014) 696–709.

[16] L. Lee, W. E. L. Grimson, Gait analysis for recognition and classification, in: Int. Conf. Automatic Face and Gesture Recognit., IEEE, 2002, pp. 155–162.

[17] L. Wang, H. Ning, T. Tan, W. Hu, Fusion of static and dynamic body biometrics for gait recognition, IEEE Trans. Circuits Syst. Video Technol. 14 (2) (2004) 149–158.

[18] Y. Chai, et al., A novel human gait recognition method by segmenting and extracting the region variance feature, in: Proc. Int. Conf. Pattern Recognit. (ICPR), Vol. 4, 2006, pp. 425–428.

[19] D. Cunado, M. S. Nixon, J. N. Carter, Automatic extraction and description of human gait models for recognition purposes, Computer Vision and Image Understanding 90 (1) (2003) 1–41.

[20] G. Ariyanto, M. S. Nixon, Marionette mass-spring model for 3d gait biometrics, in: Int. Conf. Biometrics, IEEE, 2012, pp. 354–359.

[21] W. Zeng, C. Wang, F. Yang, Silhouette-based gait recognition via deterministic learning, Pattern Recognit. 47 (11) (2014) 3568–3584.

[22] M. H. Khan, Human Activity Analysis in Visual Surveillance and Healthcare, Vol. 45, Logos Verlag Berlin GmbH, 2018.

[23] Y. Makihara, T. Tanoue, D. Muramatsu, Y. Yagi, S. Mori, Y. Utsumi, M. Iwamura, K. Kise, Individuality-preserving silhouette extraction for gait recognition, IPSJ Trans. on Comput. Vis. Appl. 7 (2015) 74–78.

[24] H. Yang, S. Lee, Reconstruction of 3d human body pose for gait recognition, in: Int. Conf. Biometrics, Springer, 2006, pp. 619–625.

[25] J. Portillo-Portillo, et al., Cross view gait recognition using joint-direct linear discriminant analysis, Sensors 17 (1) (2016) 6.

[26] A. Kale, A. R. Chowdhury, R. Chellappa, Towards a view invariant gait recognition algorithm, in: Proc. Conf. Advanced Video and Signal Based Surveillance, IEEE, 2003, pp. 143–150.

[27] F. Jean, R. Bergevin, A. Albu, Computing and evaluating view-normalized body part trajectories, Image Vis. Comput. 27 (9) (2009) 1272–1284.

[28] N. Liu, J. Lu, Y. Tan, Joint subspace learning for view-invariant gait recognition, IEEE Signal Process. Lett. 18 (7) (2011) 431–434.

[29] Z. Zhang, et al., Gii representation-based cross-view gait recognition by discriminative projection with list-wise constraints, IEEE Trans. Cybern.

[30] R. Martín-Félez, T. Xiang, Gait recognition by ranking, in: ECCV, Springer, 2012, pp. 328–341.

[31] R. Bodor, et al., View-independent human motion classification using image-based reconstruction, Int. J. Comput. Vis. 27 (8) (2009) 1194–1206.

[32] Z. Zhang, N. F. Troje, View-independent person identification from human gait, Neurocomputing 69 (1-3) (2005) 250–256.

[33] G. Zhao, et al., 3d gait recognition using multiple cameras, in: Int. Conf. Automatic Face and Gesture Recognit., IEEE, 2006, pp. 529–534.

[34] M. Hu, et al., View-invariant discriminative projection for multi-view gait-based human identification, IEEE Trans. Inf. Forensics Security 8 (12) (2013) 2034–2045.

[35] K. Bashir, T. Xiang, S. Gong, Cross view gait recognition using correlation strength, in: BMVC, 2010, pp. 1–11.

[36] W. Kusakunniran, et al., Multiple views gait recognition using view transformation model based on optimized gait energy image, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), IEEE, 2009, pp. 1058–1064.

[37] W. Kusakunniran, et al., Support vector regression for multi-view gait recognition based on local motion feature selection, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), IEEE, 2010, pp. 974–981.

[38] Y. Makihara, et al., Gait recognition using a view transformation model in the frequency domain, in: ECCV, Springer, 2006, pp. 151–163.

[39] R.-E. Fan, et al., Liblinear: A library for large linear classification, J. Mach. Learn. Res 9 (Aug) (2008) 1871–1874.

[40] S. Zheng, et al., Robust view transformation model for gait recognition, in: Proc. Int. Conf. Image Process. (ICIP), IEEE, 2011, pp. 2073–2076.

[41] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Cross-view and multi-view gait recognitions based on view transformation model using multi-layer perceptron, Pattern Recognit. Lett. 33 (7) (2012) 882–889.

[42] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: Int. Conf. Pattern Recognition. (ICPR), Vol. 4, IEEE, 2006, pp. 441–444.

[43] H. Iwama, et al., 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) (2012) 1511–1521.

[44] D. Tan, et al., Uniprojective features for gait recognition, in: Proc. Int. Conf. Biom., Springer, 2007, pp. 673–682.

[45] F. M. Castro, M. J. Marín-Jiménez, N. Guil, N. P. de la Blanca, Multimodal feature fusion for CNN-based gait recognition: an empirical comparison, arXiv preprint arXiv:1806.07753.

[46] C. Song, Y. Huang, Y. Huang, N. Jia, L. Wang, Gaitnet: An end-to-end network for gait based human identification, Pattern Recognit. 96 (2019) 106988.

[47] M. H. Khan, M. S. Farid, M. Grzegorzek, Spatiotemporal features of human motion for gait recognition, Signal Image Video Process. 13 (2) (2019) 369–377.

[48] M. H. Khan, M. S. Farid, M. Grzegorzek, Using a generic model for codebook-based gait recognition algorithms, in: Int. Workshop Biometrics Forensics (IWBF), IEEE, 2018, pp. 1–7.

[49] M. H. Khan, M. S. Farid, M. Grzegorzek, A generic codebook based approach for gait recognition, Multimed. Tools Appl. (2019) 1–24.

[50] M. J. Marín-Jiménez, F. M. Castro, Á. Carmona-Poyato, N. Guil, On how to improve tracklet-based gait recognition systems, Pattern Recognit. Lett. 68 (2015) 103–110.

[51] W. Gong, M. Sapienza, F. Cuzzolin, Fisher tensor decomposition for unconstrained gait recognition, Training 2 (3).

[52] A. Iosifidis, A. Tefas, I. Pitas, Activity-based person identification using fuzzy representation and discriminant learning, IEEE Trans. Inf. Forensics Security 7 (2) (2011) 530–542.

[53] A. Iosifidis, A. Tefas, I. Pitas, Person identification from actions based on artificial neural networks, in: IEEE Symp. Computational Intell. Biometrics Identity Manag. (CIBIM), IEEE, 2013, pp. 7–13.

[54] A. Iosifidis, N. Nikolaidis, I. Pitas, Movement recognition exploiting multi-view information, in: Proc. Int. Workshop Multimed. Signal Process. (MMSP), IEEE, 2010, pp. 427–431.

[55] O. Chapelle, S. S. Keerthi, Efficient algorithms for ranking with svms, Information retrieval 13 (3) (2010) 201–215.

[56] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, W. Meng, Coupled bilinear discriminant projection for cross-view gait recognition, IEEE Trans. Circuits Syst. Video Technol.

[57] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, W. Meng, Coupled patch alignment for matching cross-view gaits, IEEE Trans. Image Process. 28 (6) (2019) 3142–3157.

[58] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, W. Meng, A general tensor representation framework for cross-view gait recognition, Pattern Recognit. 90 (2019) 87–98.

[59] K. Shiraga, et al., Geinet: View-invariant gait recognition using a convolutional neural network, in: Int. Conf. Biometrics, IEEE, 2016, pp. 1–8.

[60] S. Yu, H. Chen, E. B. G. Reyes, N. Poh, Gaitgan: Invariant gait feature extraction using generative adversarial networks., in: CVPR Workshops, 2017, pp. 532–539.

[61] Y. Zhang, Y. Huang, S. Yu, L. Wang, Cross-view gait recognition by discriminative feature learning, IEEE Trans. Image Process. 29 (2019) 1001–1015.

[62] H. Chao, Y. He, J. Zhang, J. Feng, Gaitset: Regarding gait as a set for cross-view gait recognition, in: Proc. of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8126–8133.

[63] M. H. Khan, M. S. Farid, M. Zahoor, M. Grzegorzek, Cross-view gait recognition using non-linear view transformations of spatiotemporal features, in: Proc. Int. Conf. Image Process. (ICIP), IEEE, 2018, pp. 773–777.

[64] Laptev, Lindeberg, Space-time interest points, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2003, pp. 432–439 vol.1. doi:10.1109/ICCV.2003.1238378.

[65] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comput. Vis. 103 (1) (2013) 60–79.

[66] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, 2005.

[67] M. H. Khan, M. S. Farid, M. Grzegorzek, Person identification using spatiotemporal motion characteristics, in: Proc. Int. Conf. Image Process. (ICIP), IEEE, 2017, pp. 166–170.

[68] J. Sánchez, et al., Image classification with the fisher vector: Theory and practice, Int. J. Comput. Vis. 105 (3) (2013) 222–245.

[69] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, Comput. Vis. Image Underst. 150 (2016) 109–125.

[70] H. Rahmani, A. Mian, M. Shah, Learning a deep model for human action recognition from novel viewpoints, IEEE Trans. Pattern Anal. Mach. Intell.

[71] A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. ICML, Vol. 30, 2013, p. 3.

[72] Y. Bengio, Practical recommendations for gradient-based training of deep architectures, in: Neural networks: Tricks of the trade, Springer, 2012, pp. 437–478.

[73] R. Girshick, Fast r-CNN, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 1440–1448.

[74] F. Chollet, et al., Keras (2015).

[75] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning., in: OSDI, Vol. 16, 2016, pp. 265–283.

[76] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural networks for machine learning 4 (2) (2012) 26–31.

[77] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, Y. Bengio, An empirical evaluation of deep architectures on problems with many factors of variation, in: Proc. ICML, ACM, 2007, pp. 473–480.

[78] A. Mansur, Y. Makihara, D. Muramatsu, Y. Yagi, Cross-view gait recognition using view-dependent discriminative analysis, in: Proc. Int. Joint Conf. on Biometrics (IJCB), IEEE, 2014, pp. 1–8.

[79] I. Rida, X. Jiang, G. L. Marcialis, Human body part selection by group lasso of motion for model-free gait recognition, IEEE Signal Process. Lett. 23 (1) (2016) 154–158.

[80] C. Li, et al., Deepgait: a learning deep convolutional representation for view-invariant gait recognition using joint bayesian, Applied Sciences 7 (3) (2017) 210.

[81] Q. Chen, Y. Wang, Z. Liu, Q. Liu, D. Huang, Feature map pooling for cross-view gait recognition based on silhouette sequence images, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), 2017, pp. 54–61. doi:10.1109/BTAS.2017.8272682.

[82] D. Muramatsu, Y. Makihara, Y. Yagi, View transformation model incorporating quality measures for cross-view gait recognition, IEEE Trans. Cybern. 46 (7) (2016) 1602–1615.

[83] D. Muramatsu, et al., Gait-based person recognition using arbitrary view transformation model, IEEE Trans. Image Process. 24 (1) (2015) 140–154.