CROSS-VIEW GAIT RECOGNITION USING NON-LINEAR VIEW TRANSFORMATIONS OF SPATIOTEMPORAL FEATURES

Muhammad Hassan Khan[†], *Muhammad Shahid Farid*[‡], *Maryiam Zahoor*[‡], *Marcin Grzegorzek*[†,*]

[†]Research Group of Pattern Recognition, University of Siegen, Siegen, Germany
[‡]College of Information Technology, University of the Punjab, Pakistan
[*]University of Economics in Katowice, Katowice, Poland

ABSTRACT

This paper presents a novel cross-view gait recognition technique based on the spatiotemporal characteristics of human motion. We propose a deep fully-connected neural network with unsupervised learning which transfers the gait descriptors from multiple views to the single canonical view. The proposed non-linear network learns a single model for all videos captured from different viewpoints and finds a shared high-level virtual path to map them on a single canonical view. Therefore, the model does not require any labels or viewpoint information in the learning phase. The network is learned only once using the spatiotemporal motion features of the gait sequences from several viewpoints, later it is used to construct the cross-view gait descriptors for the gallery and the probe sets. The descriptors are classified using simple linear support vector machine. Experiments carried out on the benchmark cross-view gait dataset, CASIA-B, and comparisons with the state-of-the-art demonstrate that the proposed method outperforms the existing cross-view gait recognition algorithms.

Index Terms— Cross-view gait recognition; spatiotemporal gait features; view transformation

1. INTRODUCTION

Gait recognition refers to the problem of identifying an individual using his/her walking style. Unlike other biometrics such as fingerprints, facial features, earlobes; it does not require human interaction with the imaging system and can be collected at low resolution in a non-invasive and hidden manner. Although gait is not as powerful as other biometric modalities, its characteristic to recognize an individual from a distance and without any interaction makes it irreplaceable in many applications including visual surveillance. Gait recognition, however, is challenging as factors like clothing, shoes, walking surface, injuries and viewpoint can affect the gait. Among these, the appearance changes in a person's walk due to change in viewpoint is the most challenging and it is unavoidable too in the practical surveillance systems. It introduces intra-class variations which are always larger than inter-class variations caused by other covariates [1].

Numerous cross-view gait recognition techniques have been proposed lately which can be categorized into three groups: view-invariant gait features, construction of 3D gait descriptors, and view transformation models. The approaches placed in the first category construct a view invariant gait descriptor by transforming the gait sequences of different views into a common space [2–4]. In [2], a perspective projection model is proposed to obtain side-view gait images from an arbitrary viewpoint by assuming that the individual is a 2D planar object in the sagittal plane. The authors in [3] proposed the transformation of motion trajectories from an arbitrary view to a standard plane and their similarities are computed to identify the individuals. The algorithm presented in [5] employed subspace learning and used direct linear discriminant analysis to create a single projection model for classification. Metric learning based approaches compute a weighting vector comprising the similarity score related to each feature, which is used to estimate the recognition score [4]. These approaches perform good in specific scenarios, they are hard to generalize for other cases and their feature extraction phase is also disrupted due to self occlusion [1].

3D gait descriptors computed using multiple calibrated cameras are another mean to recognize individuals in multi-view environment. The approaches in this category construct a 3D gait model of an individual using the 2D gait information of an arbitrary view. The authors in [6] proposed a 3D visual hull model to construct the gait features using the input from multiple cameras. In [7], a 3D linear model is proposed to construct view-independent gait features using Bayesian rules. Such techniques, however, require expensive setup of multiple calibrated cameras and huge computation. Moreover, they can only be used in a controlled environment and therefore they are not suitable for real-world applications [5].

View transformation model (VTM) has emerged as a popular mean to achieve cross-view gait recognition. These

This research was partially supported by the University of the Punjab and the German Federal Ministry of Education and Research within the project "SenseVojta: Sensor-based Diagnosis, Therapy and Aftercare According to the Vojta Principle" (Grant Number: 13GW0166E).

approaches learn a mapping or transformation relationship among the gait features as perceived from different viewpoints. A singular value decomposition based VTM to project the gait features from one view to another is proposed in [8]. The algorithm in [9] computes the local motion gait features and builds a VTM using support vector regression. In [10,11], the canonical correlation analysis (CCA) is used to project each pair of gait sequence into two subspaces with maximal correlation. Hu et al. [12] proposed enhanced Gabor gait (EGG) feature which uses a non-linear mapping to encode the statistical and structural characteristics of gait across the views. In [13], a unitary linear projection is proposed to construct a cross-view gait descriptor. Wu et. al [1] proposed a deep convolutional neural network (CNN) to measure the similarity between the gait features of different views. Compared to other approaches, the VTM based algorithms have shown excellent recognition accuracies. However, the majority of these techniques learn multiple mapping matrices usually one for each pair of viewpoints.

Contrary to the existing techniques, the proposed deep neural network (DNN) based method learns a single model to transfer the gait characteristics from multiple viewpoints to the single canonical view. Our learning scheme is based on the observation that the gait characteristics of a person from different viewpoints still have a common structure that makes it different from others. Therefore, the gait related features should be separated from the viewpoint related features which is not linearly possible. In particular, the proposed algorithm exploits the spatiotemporal features of human walk and trains a DNN in an unsupervised manner. The network indeed finds a virtual path to map the gait descriptors from different views to the single canonical view. It is worth noticing that we learn a single model for this mapping using a pretty small and unlabeled set of multiview walking sequences. Therefore, the labels are neither required in the learning of our network nor in the construction of cross-view gait descriptors. Finally, the learned model is used to transform the gallery and the probe gait sequences which are fed to the subsequent classifier with their respective labels for classification. The experimental evaluation on a large multiview gait dataset, CASIA-B [14], and comparisons with existing methods confirm the superior performance of the proposed method.

2. PROPOSE METHOD

The proposed algorithm works in three steps. First, the spatiotemporal based gait representation is computed from the gait video sequences. Second, we train a DNN in unsupervised mode which finds a shared high-level virtual path to map the gait descriptors from different views to the single canonical view. Third, the gallery and the probe sequences are transformed using the trained network and fed to the subsequent classifier with their respective labels. We used the simple linear support vector machine (SVM) as classifier to demonstrate the effectiveness of our model.

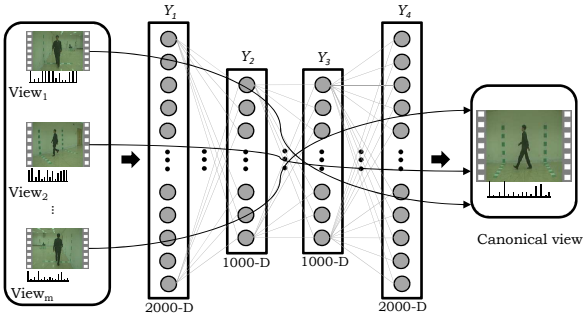

Fig. 1. Architecture of the proposed deep neural network.

2.1. Spatiotemporal based Gait Representation

Most existing gait descriptors operate on the sequences of extracted binary silhouettes of human. Therefore, their performance depends upon the variations in the silhouette shapes and also on the silhouette segmentation accuracy. An inaccurate segmentation may result in poor recognition accuracy [15]. In our earlier work [16], we proposed a spatiotemporal based gait representation which neither requires the silhouette segmentation nor the gait-cycle estimation. Local motion descriptors such as motion boundary histogram (MBH) and histogram of oriented gradient (HOG) were computed from the gait videos along with the dense trajectories. The HOG and MBH local descriptors are encoded using Fisher vector [17] encoding and a codebook based on Gaussian mixture model (GMM). The derived features have demonstrated the best performance in single-view gait recognition [16], and therefore the same gait representation is chosen to construct a cross-view gait descriptor. The length of each spatiotemporal gait feature is set to $2,000$ using the principal component analysis (PCA).

2.2. View Transfer Model

The proposed algorithm learns a DNN to transfer the gait descriptors of different viewpoints to the single canonical view. Most existing VTM based approaches [7, 11–13] use a set of linear transformations between the source and the target viewpoints to obtain this mapping. Therefore, such approaches are unable to capture the non-linear manifolds where the realistic gait scenarios lie [1]. To overcome this problem, we propose a non-linear deep network that learns a shared high-level virtual path to map the gait sequences from all source viewpoints to the same canonical view. Our learning scheme is based on two observations: first, the gait descriptors from different viewpoints posses the same high-level representation that makes it unique from others. Second, there are m different virtual paths connecting the m source viewpoints to the canonical view (Fig. 1), and the proposed network aims to

force them to learn a single non-linear virtual path. Thus, the source viewpoints are mapped to the canonical view through intermediate virtual views along the non-linear path.

The proposed DNN consists of 4 layers with l_h units in each layer, where l is representing the layer number and h is the hidden units per layer. During training, the output of each layer is passed as input to the next layer.

$$Y_1 = f(W_1 x_{ij} + b_1), \tag{1}$$

where Y_1 denotes the output of the first layer, x_{ij} is the j-th training instance of the i-th viewpoint, and $f(.)$ is the non-linear activation function. Moreover, W_1 is the weight matrix and b_1 is the bias vector to be learned for the first layer. We used the leakyReLU (Leaky Rectified Linear Unit) as the activation function. It is a variant of ReLU and exhibits better results. Unlike other activation functions *e.g*. sigmoid, it does not suffer from the vanishing gradient problem [1,18]. In contrast to ReLU, it allows a small non-zero gradient when the unit is saturated and not active [18]. Specifically, it assigns a small slope to the negative part instead of dropping it:

$$f(x) = \begin{cases} x & \text{if } x > 0; \\ ax & \text{otherwise} \end{cases} \tag{2}$$

where α is a small constant. Thus, for a given training instance $x \in \Re$, the output of the first layer Y_1 is forwarded as an input to the second layer, and so on. The output of the last fully connected layer can be computed as,

$$t(x_{ij}) = Y_l = f(W_l Y_{l-1} + b_l), \tag{3}$$

where $t(x_{ij})$ represents the non-linear transformation of x_{ij} using W_l and b_l. The number of units in each layer are empirically chosen to ensure that the network can efficiently learn the underlying structure of the data. The size of the hidden units in the first and the last layers is set to $2,000$ due to the dimension of our spatiotemporal gait features (Section 2.1), and the size each intermediate layer is set to $1,000$. Thus, the redundant information in the input features can be removed by mapping them to a high-level but low-dimensional representation and this computation is performed in the first two connected layers. Later, this low-dimensional compact representation is mapped back to the high-dimensional output layer using the last two connected layers as shown in Fig. 1. The learning of the proposed network consists of minimizing the loss function of the reconstruction error over all training samples from all viewpoints by updating a set of parameters $\Theta = \{W_l, b_l; l = 1, 2, ..., 4\}$. Moreover, we add weight decay J_w to penalize the objective function to reduce the effect of over-fitting. For m viewpoints and n instances in each view, the reconstruction error e_Θ is defined as,

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

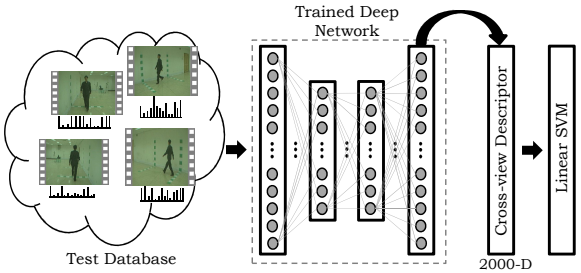

Fig. 2. Construction of the cross-view descriptor and classification using linear SVM. The output of the last layer is used as cross-view gait descriptor.

where c is the canonical view and λ is the weight decay parameter. Large weights may cause highly curved and non-smooth mappings [19]. The weight decay penalizes the large weights and keeps them small to make the mappings smooth and to reduce the over-fitting. We used mini-batch stochastic gradient descent method through back-propagation to minimize the objective function over all the training samples. During training, the input to the first and the last layers of our network are the spatiotemporal gait features from the source viewpoints (*i.e*. all possible viewpoints) and the canonical view (*i.e*. $90°$), respectively. Thus, the proposed network tries to get the output of the descriptor to be close to its canonical view regardless of its input view.

2.3. Cross-view Gait Descriptor and Classification

It can be observed from (3) that the non-linear transformation function $t(x_{ij})$ provides the canonical view representation of gait sequences obtained from any unknown viewpoint, which can be used as cross-view gait representation. Since the proposed DNN consists of a set of non-linear transformations $\{Y_1, \ldots, Y_4\}$ to map the gait sequences from different viewpoints to a canonical view, the output of the last layer in the proposed DNN is selected as cross-view gait descriptor (Fig. 2) because it encodes the influence of all these transformations and provides the canonical-view representation of gait. For classification, the gait descriptors of gallery set are selected and their cross-view gait representation is obtained using the learned network. They are used to train a classifier along-with their respective labels. We used a simple linear SVM [20] to show the strength of our proposed deep network. In testing, the cross-view descriptors of the probe sequences are computed and fed to SVM to identify the walker.

3. EXPERIMENTS AND RESULTS

The performance of the proposed algorithm is evaluated on the benchmark CASIA-B gait dataset and the results are compared with the current state-of-the-art.

Table 1. Performance on CASIA-B gait dataset. The gallery set (θ_g) consists of all views excluding the view in the probe set (θ_p). 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 [9]	-	28	29	34	35	44	45
TSVD [23]	-	39	33	42	49	50	54
CMCC [11]	46.3	52.4	48.3	56.9	-	-	-
ViDP [13]	-	59.1	50.2	57.5	83.5	76.7	80.7
CNN [1]	54.8	77.8	64.9	76.1	90.8	85.8	90.4
Proposed	**58.5**	**97.5**	**91.5**	**97**	**98.5**	**93.5**	**98.0**

Table 2. Performance on CASIA-B gait dataset with gallery view 90°. The best results are marked in bold.

θ_p	0°	18°	36°	54°	72°	108°	126°	144°	162°	180°
JDLDA [5]	20	25	37	58	94	-	-	-	-	-
Method [25]	12	20	30	60	92	92	62	35	19	12
MvDA [26]	17	27	36	64	95	-	-	-	-	-
GII [27]	17	26	54	84	98	98	84	50	25	14
JSL [4]	20.5	35.5	56.5	81.5	96.5	**96**	**89.5**	50	34.5	21.5
DATER [12]	3.2	7.4	16.8	48.1	66.5	-	-	-	-	-
Method [11]	18	24	41	66	96	95	68	41	21	13
Proposed	**51.5**	**53**	**63**	**85.5**	**99**	83	55.5	**54**	**52**	**51**

3.1. Implementation Details

The deep network is constructed using an open source library Keras [21] with Tensorflow [22] at back-end, and trained using back-propagation with logistic regression loss defined in (4). The network parameters Θ are initialized using simple random initialization method [19]. All the bias terms b_l are initialized with zero and the weight matrix W_l is initialized using the Gaussian distribution with zero mean and 0.05 standard deviation. The size of the mini-batch is set to 64 and the values of λ and the learning-rate are set to 0.0001 and 0.001, respectively, using multi-resolution search [19].

3.2. Recognition Results on CASIA-B

The CASIA-B dataset comprises the gait sequences of 124 subjects captured using eleven different viewpoints: 0°, 18°, 36°, . . ., 180°. Ten walk sequences are recorded for each subject with three variations: normal walk (nm), walk with bag (bg), and walk with coat (cl). Among these ten sequences, six belong to nm, two to each bg and cl. Cross-view gait recognition on this dataset is quite challenging due to large cross-view angles and it is even more difficult when the probe and the gallery sets belong to different walking scenarios [1, 5]. Similar to [1], the gait sequences of randomly selected 24 subjects are used to train the proposed DNN and the videos of remaining 100 subjects are used in the performance evaluation. In all experiments, the first four normal walk sequences ($nm_1 - nm_4$) of test dataset are used to form the gallery and the rest are used in different probe sets. Similar to the state-of-the-art [1,5,24], three types of experiments are performed. In the first set of experiments, the gallery set contains the gait sequences from multiple viewpoints excluding the view in the probe set. The recognition accuracies are presented in Table 1 which show that our method outperforms the compared methods in all experiments with significant margins.

In the second set of experiments which are derived from [4, 11], the gallery set contains the gait sequences from 90° viewpoint and the rest of the viewpoints are used in probe sets (θ_p), separately. The recognition accuracies are listed in Table 2 which reveal that our method outperforms the others in all experiments except two viewing angles 108° and 126° where JSL [4] performs better than our method. In the last set of experiments the robustness of the propose method is tested under various conditions on the gait sequences recorded at 54°, 90° and 126°. Similar to [24], the gallery set comprises the gait sequences of normal walk from viewing angle 36° to 144°. The results are presented in Table 3 and the statistics show that our method outperforms the compared methods in most experiments. The results of the experimental evaluation presented in Tables 1 to 3 reveal that the proposed method performed consistently better than the state-of-the-art cross-view gait recognition techniques.

Table 3. Performance on CASIA-B gait dataset under various scenarios of walk. In each block, two recognition results of bg and cl are presented, separated with '/'. The best results are marked in bold.

θ_p	54° (bg/cl)	90° (bg/cl)	126° (bg/cl)
Method [24]	94.2 / 93.5	**92.3 / 92**	95.1 / **94.2**
Method [28]	76.4 / 87.9	73.7 / 91.1	76.9 / 86.2
RLTDA [12]	80.8 / 69.4	76.5 / 72.1	72.3 / 64.4
Robust VTM [29]	40.7 / 35.4	58.2 / 50.3	59.4 / 61.3
FT-SVD [8]	26.5 / 19.8	33.1 / 20.6	38.6 / 32
CNN [1]	92.7 / 49.7	88.9 / 75.6	86.0 / 51.4
Proposed	**96 / 94.5**	87.5 / 91	**98.5** / 94

4. CONCLUSION

In this paper, we have presented a novel cross-view gait recognition method using a non-linear deep neural network with unsupervised learning. The network is trained using the spatiotemporal motion features of human gait. It transfers the gait descriptors of an individual from multiple viewpoints to a single canonical view. The cross-view gait representation of a testing instance is achieved by its forward propagation through the trained network. Excellent classification results using a simple linear SVM reveal the effectiveness of the proposed network. In the experimental evaluation carried out on a large benchmark gait dataset, our method outperformed the exiting techniques in most experiments.

5. REFERENCES

[1] Z. Wu et al., "A comprehensive study on cross-view gait based human identification with deep cnns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, 2017.

[2] A. Kale et al., "Towards a view invariant gait recognition algorithm," in *Proc. Advanced Video and Signal Based Surveillance*. IEEE, 2003, pp. 143–150.

[3] F. Jean, R. Bergevin, and A. Albu, "Computing and evaluating view-normalized body part trajectories," *Image Vis. Comput.*, vol. 27, no. 9, pp. 1272–1284, 2009.

[4] N. Liu, J. Lu, and Y. Tan, "Joint subspace learning for view-invariant gait recognition," *IEEE Signal Process. Lett.*, vol. 18, no. 7, pp. 431–434, 2011.

[5] J. Portillo-Portillo et al., "Cross view gait recognition using joint-direct linear discriminant analysis," *Sensors*, vol. 17, no. 1, pp. 6, 2016.

[6] R. Bodor et al., "View-independent human motion classification using image-based reconstruction," *Image Vis. Comput.*, vol. 27, no. 8, pp. 1194–1206, 2009.

[7] Z. Zhang and N. F Troje, "View-independent person identification from human gait," *Neurocomputing*, vol. 69, no. 1-3, pp. 250–256, 2005.

[8] Y. Makihara et al., "Gait recognition using a view transformation model in the frequency domain," in *Proc. ECCV*, 2006, pp. 151–163.

[9] 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)*, 2010, pp. 974–981.

[10] K. Bashir et al., "Cross view gait recognition using correlation strength," in *BMVC*, 2010, pp. 1–11.

[11] W. Kusakunniran et al., "Recognizing gaits across views through correlated motion co-clustering," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 696–709, 2014.

[12] H. Hu, "Enhanced gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1274–1286, 2013.

[13] M. Hu et al., "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2034–2045, 2013.

[14] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *ICPR*, 2006, vol. 4, pp. 441–444.

[15] M. H. Khan et al., "Gait recognition using motion trajectory analysis," in *Proc. Int. Conf. Comput. Recognit. Systems (CORES)*. Springer, 2017, pp. 73–82.

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

[17] J. Sánchez et al., "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[18] A. L Maas, A. Y Hannun, and A. Y Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML*, 2013, vol. 30, p. 3.

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

[20] R-E. Fan et al., "Liblinear: A library for large linear classification," *J. Mach. Learn. Res*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[21] François Chollet et al., "Keras," 2015.

[22] Martín Abadi et al., "Tensorflow: A system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[23] 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)*, 2009, pp. 1058–1064.

[24] J. Tang et al., "Robust arbitrary-view gait recognition based on 3d partial similarity matching," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 7–22, 2017.

[25] W. Kusakunniran et al., "Cross-view and multi-view gait recognitions based on view transformation model using multi-layer perceptron," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 882–889, 2012.

[26] A. Mansur et al., "Cross-view gait recognition using view-dependent discriminative analysis," in *Proc. Int. Joint Conf. on Biometrics (IJCB)*. IEEE, 2014, pp. 1–8.

[27] Z. Zhang et al., "Gii representation-based cross-view gait recognition by discriminative projection with list-wise constraints," *IEEE Trans. Cybern.*, 2017.

[28] I. Rida, X. Jiang, and G. L. Marcialis, "Human body part selection by group lasso of motion for model-free gait recognition," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 154–158, 2016.

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