Learning Optimised Representations for View-invariant Gait Recognition

Ning Jia, Victor Sanchez
Department of Computer Science
University of Warwick
{n.jia, v.f.sanchez-silva}@warwick.ac.uk

Chang-Tsun Li
School of Computing and Mathematics
Charles Sturt University
chli@csu.edu.au

Abstract

Gait recognition can be performed without subject cooperation under harsh conditions, thus it is an important tool in forensic gait analysis, security control, and other commercial applications. One critical issue that prevents gait recognition systems from being widely accepted is the performance drop when the camera viewpoint varies between the registered templates and the query data. In this paper, we explore the potential of combining feature optimisers and representations learned by convolutional neural networks (CNN) to achieve efficient view-invariant gait recognition. The experimental results indicate that CNN learns highly discriminative representations across moderate view variations, and these representations can be further improved using view-invariant feature selectors, achieving a high matching accuracy across views.

1. Introduction

The non-invasive and unobtrusive nature of gait recognition has attracted increasing attention from researchers in the field of biometrics and computer vision. On the one hand, the proliferation of CCTV cameras in public spaces and the extensive deployment of biometric systems have entailed the implementation of multi-biometrics systems, where human tracking and recognition using gait are efficient tools for narrowing down search and boosting processing speed [9]. On the other hand, when data depicting face, iris, fingerprint, or other biometric traits are inaccessible or severely degraded due to occlusion or other environmental factors, data depicting gait may still be reliable due to its higher tolerance to noise [18]. As mentioned in [2] [13], pioneering experiments have been successfully carried out to test gait biometrics on forensic tasks with real crime scene videos, and gait-based evidence has already been introduced by courts to increase the confidence of identity. The recent development of machine learning algorithms, especially representation learning techniques, has notably promoted the accuracy of gait recognition systems. As a result, the potential of gait recognition for security and commercial purposes has increased considerably.

Gait recognition approaches can be broadly classified into two categories: model-based and appearance-based. Model-based gait recognition refers to identifying people by modelling their distinctive gait characteristics with underlying mathematical structures [3]. It relies on high quality gait sequences captured under controlled environments (e.g., indoor environments, close-distance between subject and camera, multi-view cameras, in-depth cameras or kinetic cameras). Unfortunately, the associated restrictions to specific sensors and low tolerance to low quality data, make model-based methods less applicable to outdoor gait recognition. Appearance-based methods adopt human profiles, i.e., gait silhouettes, as the source of features to build effective gait templates. A commonly used appearance-based template is the Gait Energy Image (GEI), which averages all the binary silhouettes from a gait cycle and generates a single gait template [6]. Experiments on large-scale gait datasets suggest that GEIs are among the most statistically stable and efficient templates for gait recognition [8] [16]. Apart from their high efficiency, the main advantage of appearance-based approaches is the ease of access to gait representations. In this paper, we thus adopt GEIs as the gait representation templates.

Despite its potential, gait is still not widely used as forensic evidence by courts or law enforcement, due to the fact that gait features may be easily contaminated by a number of factors, e.g., health conditions of the subject, time elapse, clothes, and carrying conditions. Despite two decades of research and development, hard challenges remain unsolved, including large viewpoint variations between the templates and query data. In order to achieve view-invariant gait recognition, we propose to learn gait representations by using a convolutional neural network (CNN) [14], followed by a representation optimisation technique to improve the matching accuracy across all views. In our previous work [10], we have proposed a view-invariant feature selector (ViFS) to automate the feature selection process and perform fast and efficient view-invariant gait recognition. ViFS is designed to select the most representative features from a multi-view gallery dataset when matching probe data captured at a specific viewpoint. In this paper, we further enhance the performance of the ViFS-based framework with representations learned by a well-trained CNN. We notice that the used in conjunction with feature enhances, e.g., Linear Discriminant Analysis (LDA), the performance of the ViFS-based framework is correlated with the cross-view feature learning ability of the feature enhancers [10]. Since CNN has been shown to have very strong feature learning capabilities [4] [11] [22], we use CNN as a feature enhancer of ViFS for view-invariant gait recognition. Specifically, we first train a CNN with sufficient multi-view data. We then use the CNN as the feature enhancer to obtain multi-view gait feature maps from a specific layer of the network. ViFS is finally used to reconstruct gallery features from the multi-view gallery feature maps to match with the probe features. The joint power of CNN feature maps and ViFS achieves a very high recognition accuracy on the CASIA Dataset B across 11 views with gallery data from only 2 views, which suggests the great potential of the proposed framework.

The rest of this paper is organized as follows. Section 2 reviews recent works related to view-invariant gait recognition and the implementation of CNNs for gait recognition. Section 3 explains the proposed framework that merges representation learning and ViFS. Section 4 details the experimental setting and discusses the results. Section 5 concludes this paper and provides ideas for future work.

2. Related Works

Since gait sequences are normally acquired from a distance at low resolutions with occlusions, it is difficult to extract model-based parameters (height, length of limbs, joint angle, etc.) from the captured gait sequences. Therefore most researchers adopt appearance-based features. Appearance-based view-invariant gait recognition algorithms can be classified into two categories: 1) those based on view-invariant features, and 2) those based on unitary projections. In the first category, researchers seek for view-invariant features from single-view gait silhouette sequences, and perform recognition under a lateral view (90°). For example, Kusakunniran et al. propose the Gait Texture Image (GTI) and apply transform invariant low-rank textures to obtain common canonical view (later view - 90°) gait features from other view angles [12]. However, one limitation of their method is that it is difficult to transfer features from front or back views to the lateral view. Goffredo et al. propose model based view-invariant gait features, which use lower limbs' pose estimation to perform viewpoint rectification [5] . This method, however, suffers from many of the disadvantages of model-based methods.

Because there are no view-invariant features on gait silhouettes, cross-view gait matching is normally performed by means of learning projections among views. Hu et al. propose the view-invariant discriminative projection (ViDP) for cross-view gait recognition, which iteratively optimise the construction of the local affinity matrix and achieves discriminant feature projection without knowing the view angles of the probe data [7]. Makihara et al. propose a singular value decomposition (SVD)-based method to deal with camera viewpoint changes, the so-called view transformation model (VTM) [17], which has been further improved in [19, 20, 30]. View-projection based methods, however, are usually not capable of providing satisfying results under large viewpoint variations ($>= 36^\circ$).

Recently, CNNs have been introduced to tackle gait recognition challenges. Alotaibi et al. apply a full convolutional network (fullconvnet) with 4 convolutional layers and a softmax layer for simple gait recognition tasks, i.e., matching gallery and probe under the same conditions [1]. Yan et al. use a 5-layer network with 3 convolutional layers and 2 fully connected layers for gait recognition with different clothing conditions [27]. They also introduce a multi-task learning approach, which performs gait recognition, view prediction and scene prediction simultaneously. According to their report, multi-task learning can accelerate the convergence of CNNs in the training process. However, as reported in their paper, the performance of their network on challenging tasks has no significant improvement compared with that of traditional approaches. Shiraga et al. propose a 4-layer network consisting of 2 convolutional layers and 2 fully connected layers, and use it for large-scale gait recognition on the OU-ISIR Large Population Dataset [23]. Their network has major advantages over other approaches on large-scale datasets if the viewpoint variation is small (at most 30 °). The authors also show that CNN-based methods can significantly reduce the equal error rates (EER) and thus improve the gait verification accuracy. A thorough study on CNN-based gait recognition is provided by Wu et al., where they extensively evaluate the effect of the training procedure and network architecture on the performance [26]. Their experimental results show that the pair-image training network is capable of outperforming other approaches by a large margin. The feature maps learned by the CNN have strong discriminative power and good robustness for cross-condition gait recognition (e.g., different clothing or carrying conditions between gallery and probe data). However, cross-view recognition with large viewpoint variations (54°and above) is still not ideal.

When multi-view gait templates are obtained, or depth information is available, it is possible to reconstruct 3D or 2.5D models representing the human body, from which arbitrary views of gait sequences can be obtained by projection, and parameters associated with body parts can be easily measured. Tang et al. [25] propose to construct paramet-

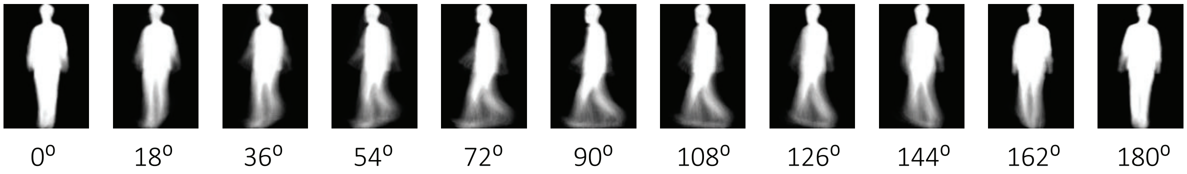

Figure 1. Example GEIs from 11 different views, computed from silhouettes from the CASIA Dataset B [28].

ric 3D gait models from three cameras and use partial similarity matching to improve recognition rates. Their method achieves promising results on several major gait datasets. Similarly, Luo et al. [15] propose to use 3D gait models and sparse representation-based classification to perform view-invariant classification. However, 3D model based approaches, including [25] and [15], require a specifically designed multi-view database for model construction and training purposes.

3. Proposed Framework

The proposed framework uses GEIs as gait features. With n silhouette images in one gait period, a GEI is defined as $\mathbf{G} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{I}_k$, where $\mathbf{I}_k$ is the kth silhouette image. Examples of GEIs are illustrated in Fig. 1.

3.1. CNN Feature Maps

In this paper, we use a 4-layer CNN to learn gait representations from the training dataset. This CNN consists of 3 convolutional layers, each followed by a rectifier and a max pooling layer, and 1 fully connected layer followed by a rectifier and dropout. The convolutional layers are designed for local feature extraction, where the neurones are locally connected to those of the previous layer and their weights are shared across all spatial locations. The rectifier layer uses the rectified linear unit (ReLU), $f(x) = \max(0, x)$ [21], for adding non-linearity to the feature maps. The pooling layer uses the max pooling method, which down-samples the feature maps while preserving the maximum value for each local region in order to attain rotation-invariant feature extraction, as well as dimensionality reduction. Dropout is designed to reduce the neurons' co-adaptation and eventually reducing overfitting during the training phase, thus improving the representation learning ability of the network [24]. As suggested by Wu et al. [26] and Isola et al. [29], using two inputs in one layer forces the network to learn the difference between a pair of features, thereby enabling the network to learn representations in a discriminative manner. We follow this idea and use a similar parameter setting as in [26] to train the CNN model. The filters in the convolutional layers are set to 7×7 with stride 1, the poolings are of size 2×2 with a stride of 2. The original size of GEIs is 128×88, which is resized to 126×126 before feeding the GEIs to the network. The size of the output feature maps after 3 convolutional layers is 11×21.

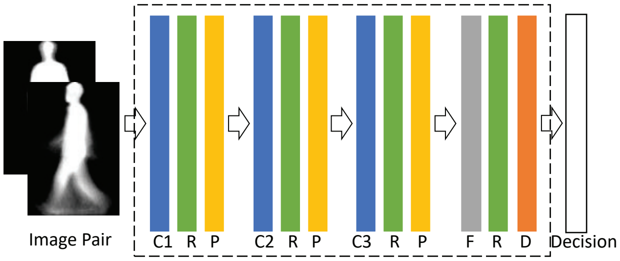

Figure 2. The network architecture for representation learning. Rectangles in different colors are used to indicate network layers. Blue (Cn): nth convolutional layer. Green (R): rectifier layer. Yellow (P): pooling layer. Grey (F): fully-connected layer. Orange (D): dropout layer. The inputs to the network are 2-channel images composed of two GEIs.

3.2. Feature Optimization

For the sake of simplicity, the learning process of the proposed CNN model f is denoted as the composition of convolutional function f_C and decision function f_D, i.e., $f = f_C \circ f_D$. f_C proceeds the feature map learning process, corresponding to layers C1 to C3 in Fig. 2, and f_D corresponds to the fully connected layer, F. Function f is trained to map input data $\mathbf{x}$ to an output vector $\mathbf{y}$, i.e., the output labels. In this paper, $\mathbf{x}$ refers to a 2-channel input image, which is composed of two grey-scale GEIs from two distinct views. The learned gait representation,$\boldsymbol{g}$, before f_D is $\boldsymbol{g} = f_C(\mathbf{x})$.

Let us assume that we have $n_{\mathcal{G}}$ gallery samples from h different views in gallery set $\mathcal{G} = \{\boldsymbol{g}_i\}_{i=1}^{n_{\mathcal{G}}}$, and one probe sample $\boldsymbol{p}$ from a specific view angle in probe set $\mathcal{P}^1$. Under the scenario that the view angles of the gallery and probe samples are unknown, our objective is to find a feature vector $\mathbf{w} = \{w_i\}_{i=1}^{h}$ that minimises the objective function [10]:

$$f(\mathbf{w}) = \|\mathcal{G}\mathbf{w}^{\intercal} - \boldsymbol{p}\|^2 = \|\sum_{i=1}^{h} w_i \boldsymbol{g}_i - \boldsymbol{p}\|^2. \quad (1)$$

After obtaining the ViFS projection basis, $\mathbf{w}$, from the train-

[1] Both $\boldsymbol{g}$ and $\boldsymbol{p}$ are the CNN feature maps.

Table 1. Cross-view matching using CNN feature maps. G: Gallery; P: Probe

P G	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
0°	100	100	92	92	68	64	68	84	88	100	100
18°	100	100	100	100	84	76	76	92	92	100	92
36°	92	100	100	100	96	84	92	92	92	92	80
54°	80	100	100	100	100	100	96	100	92	84	72
72°	56	80	100	100	100	100	100	100	100	72	60
90°	52	80	88	96	100	100	100	100	100	72	48
108°	52	76	88	96	100	100	100	100	100	72	56
126°	68	84	84	100	100	100	100	100	100	92	72
144°	84	96	96	96	96	88	100	100	100	100	76
162°	96	100	96	84	80	68	80	100	100	100	100
180°	96	96	84	80	72	64	64	96	88	100	100

ing procedure, the gallery set after feature extraction is recosntructed by $\hat{\mathcal{G}} = \mathcal{G}\mathbf{w}^{\intercal}$.

We use Euclidean distance to obtain matching scores between gallery and probe feature maps. The Euclidean distance between gallery $\hat{\mathcal{G}}$ and probe $\mathcal{P}$ is calculated as:

$$D(\hat{\mathcal{G}}_i, \mathcal{P}_l) = \|\hat{\mathcal{G}}_i - \mathcal{P}_l\|, \quad i = 1, ..., c, \tag{2}$$

where c is the number of classes. If $D(\mathcal{G}_k, \mathcal{P}_l) = \min_{i=1}^{c} D(\mathcal{G}_i, \mathcal{P}_l)$, the probe data is assigned to the same class label k of the gallery.

4. Experiments and Analysis

We now validate the cross-view performance of the prosed framework. Firstly, we present the baseline results using the CNN feature maps computed on the CASIA Dataset B. We then present the results of the framework merging CNN feature maps with ViFS.

4.1. Cross-view Matching Using CNN Feature Maps

In this experiment, we input the gallery and probe GEI templates into the network, and extract the feature maps from the penultimate layer of the CNN. We measure the Euclidean distance between the gallery and probe feature maps. The cross-view matching accuracy is tabulated in Table 1. Apart from the large view disparity cases, which are marked in grey colour, other cross-view matching results are all above 80%, suggesting that CNN feature maps have great discriminant power. By comparing these results with those reported in our previous work [10], we can conclude that CNN feature maps can then attain significant improvements compared to conventional discriminant learning methods, such as LDA.

4.2. CNN Feature Maps and ViFS

In this experiment, we apply ViFS to the multi-view gallery CNN feature maps. Let us assume that the gallery set has 2 views available, since the number of all views is 11, there are $\binom{11}{2} = \frac{11!}{2!(11-2)!} = 55$ different combinations. We select 3 representative sets for comparison. Set 1 contains the $\{0°, 90°\}$ views: the gallery views are widely spread. Set 2 contains the $\{0°, 54°\}$ views, where the 0° view attains good performance on the frontal/back views and the 54° view attains good performance on other views (18° to 144°). Set 3 contains the $\{18°, 108°\}$ views. Results of this experiement are tabulated in Table 2, where the column *Average* tabulates the average of all results across each row. As shown in this Table, Set 3 achieves very high accuracy, on average, while Set 1 and Set 2 achieve only a slight improvement compared to the average results of Table 1, which are tabulated in row *Avg*. Row *Wu el al.* tabulates the state-of-the-art CNN method presented in [26]. Their experimental setting assumes that gallery data from the 0° to the 180° views are available. Set 3 outperforms Wu et al.'s method by 3%, on average, but the results on the 72° and 90° probe data are lower than theirs.

In Table 3, we compare CNN+ViFS with two recently proposed methods by Tang et al. [25]. *Tang(9)* refers to the results of using 9 training views from 18° to 162°. *Tang(4)* refers to the results of using 4 training views, i.e., $\{36°, 72°, 108°, 144°\}$. *ViFS3(2)* refers to the proposed framework CNN+ViFS using 2 training views $\{18°, 108°\}$, i.e., Set 3. In their paper, Tang et al. only compare the results on probe data from the 18° and 162° views, thus we follow the same setting. The column *Average* of Table 3 tabulates the average of all results across each row. ViFS3(2) attains a better performance than that attained by *Tang(4)*, but with less gallery views, It performs equality well, on average, as *Tang(9)*, but with much less gallery views.

We also provide a comparison between different cases by varying the number of gallery views available, in order to explore the potential of the CNN+ViFS framework. In Case 1, we evaluate the case when the gallery templates are mainly from frontal views. In the 2-view case

Table 2. Matching results (%) using the combination of CNN feature maps and ViFS.

Set	Probe											
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Average
Set 1	100	100	92	92	80	80	80	96	92	100	100	92
Set 2	76	100	100	100	100	96	100	100	92	80	64	91.6
Set 3	100	100	100	100	92	88	96	100	100	100	92	**97.1**
Avg.	86.9	92	92.7	93.1	88	85.1	85.5	90.9	93.8	91.3	85.5	89.5
Wu et al.	88.7	95.1	98.2	96.4	94.1	91.5	93.9	97.5	98.4	95.8	85.6	94.1

Table 3. Comparisons with Tang el al.'s work. Results are in (%). [25].

Method	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Average
Tang(9)	-	94	98	99	98	99	98	98	98	93	-	97.3
Tang(4)	-	91	98	92	98	94	98	93	98	90	-	94.7
ViFS3(2)	-	100	100	100	92	88	96	100	100	100	-	**97.3**

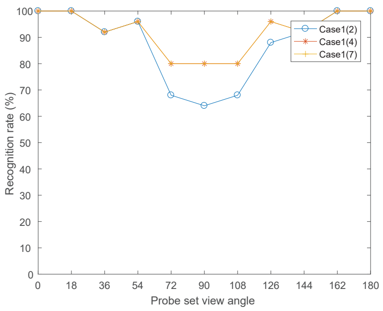

Figure 3. Recognition accuracy (%) of ViFS+CNN when gallery templates from different views are available in the gallery set. Gallery views are mainly from frontal views.

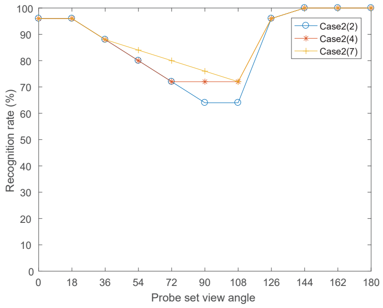

Figure 4. Recognition accuracy (%) of ViFS+CNN when gallery templates from different views are available in the gallery set. Gallery views are mainly from back views.

(Case1(2)), the gallery set contains the $\{0°, 18°\}$ views. In the 4-view case (Case1(4)), the gallery set contains the $\{0°, 18°, 36°, 54°\}$ views. In the 7-view case (Case1(7)), the gallery set contains views from 0° to 108°. As shown in Fig. 3, Case1(4) and Case1(7) attain very similar performances, which is unexpected, since more view angles are expected to improve performance. To understand these results, the corresponding ViFS projection basis, $\mathbf{w}$, are analyzed. We observe that the computed weights are almost evenly distributed and their sum is much grater than 1, which indicates that ViFS tends assign similar importance to all gallery features, regardless of the views. In Fig. 4, we observe similar (but opposite) results to those in Fig. 3, where gallery templates are mainly from back views. In the 2-view case (Case2(2)), the gallery set contains the $\{162°, 180°\}$ views. In the 4-view case (Case2(4)), the gallery set contains the $\{126°, 144°, 162°, 180°\}$ views. In the 7-view case (Case2(7)), the gallery set contains views from 72° to 180°.

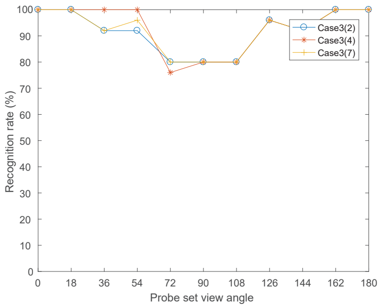

Figure 5. Recognition accuracy (%) of ViFS+CNN when gallery templates from different views are available in the gallery set. Gallery views are from widely spread frontal views.

Finally, we evaluate the case when the gallery templates are widely spread, but mainly from frontal views (see Fig. 5). In the 2-view case (Case3(2)), the gallery set contains the $\{0°, 90°\}$ views. In the 4-view case (Case3(4)), the gallery set contains the $\{0°, 18°, 54°, 90°\}$ views. In the 7-view case (Case3(7)), the gallery set contains views from 0° to 108°. The three curves in Fig. 5 are very similar, indicating that the great discriminative power of the CNN tend to *confuse* ViFS, making it difficult to select the most representative features from the multi-view feature set.

5. Conclusion

In this paper, we proposed to learn feature representations from GEIs using a 5-layer CNN, followed by ViFS to reconstruct the optimal representation from the multi-view feature set. Experimental results on CASIA Dataset B indicate that our proposed framework. which merges CNN+ViFS, outperforms existing algorithms, and has the potential to be implemented in security settings, as it performs very well with gallery data from only two distinct views. For example, the development of automatic identification systems at border control can incorporate two cameras capturing the subject whose identity is to be verified. GEIs from these two views can be fed into the proposed framework, generating reliable identification results. A fusion of face and gait would further increase the accuracy and reliability of such system.

6. Acknowledgement

This work is supported by the EU Horizon 2020 – Marie Skłodowska-Curie Actions through the project Computer Vision Enabled Multimedia Forensics and People Identification (Project No. 690907, Acronym: IDENTITY).

References

[1] M. Alotaibi and A. Mahmood. Improved gait recognition based on specialized deep convolutional neural networks. In *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7, 2015.

[2] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889, 2011.

[3] Y. Chew-Yean and M. Nixon. Model-based gait recognition. *Enclycopedia of Biometrics*, pages 633–639, 2009.

[4] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[5] M. Goffredo, I. Bouchrika, J. Carter, and M. Nixon. Self-calibrating view-invariant gait biometrics. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(4):997–1008, Aug 2010.

[6] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006.

[7] M. Hu, Y. Wang, Z. Zhang, J. Little, and D. Huang. View-invariant discriminative projection for multi-view gait-based human identification. *IEEE Transactions on Information Forensics and Security*, 8(12):2034–2045, Dec 2013.

[8] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, 7(5):1511–1521, 2012.

[9] A. K. Jain, P. Flynn, and A. A. Ross. *Handbook of Biometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[10] N. Jia, C. T. Li, V. Sanchez, and A. W. C. Liew. Fast and robust framework for view-invariant gait recognition. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, April 2017.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] W. Kusakunniran, Q. Wu, J. Zhang, Y. Ma, and H. Li. A new view-invariant feature for cross-view gait recognition. *IEEE Transactions on Information Forensics and Security*, 8(10):1642–1653, 2013.

[13] P. K. Larsen, E. B. Simonsen, and N. Lynnerup. Gait analysis in forensic medicine. *Journal of Forensic Sciences*, 53(5):1149–1153, 2008.

[14] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256, May 2010.

[15] J. Luo, J. Tang, T. Tjahjadi, and X. Xiao. Robust arbitrary view gait recognition based on parametric 3d human body reconstruction and virtual posture synthesis. *Pattern Recognition*, 2016.

[16] Y. Makihara, D. S. Matovski, M. S. Nixon, J. N. Carter, and Y. Yagi. Gait recognition: Databases, representations, and applications. *Wiley Encyclopedia of Electrical and Electronics Engineering*, 2015.

[17] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Which reference view is effective for gait identification using a view transformation model? In *Proceedings of Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 45–45, June 2006.

[18] D. S. Matovski, M. S. Nixon, S. Mahmoodi, and J. N. Carter. The effect of time on gait recognition performance. *IEEE Transactions on Information Forensics and Security*, 7(2):543–552, April 2012.

[19] D. Muramatsu, Y. Makihara, and Y. Yagi. View transformation model incorporating quality measures for cross-view gait recognition. *IEEE Transactions on Cybernetics*, PP(99):1–1, 2015.

[20] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Uddin, and Y. Yagi. Gait-based person recognition using arbitrary view transformation model. *IEEE Transactions on Image Processing*, 24(1):140–154, Jan 2015.

[21] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[22] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proceeding of International Conference on Learning Representations (ICLR)*, page 16, 2013.

[23] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *Proceedings of International Conference on Biometrics (ICB)*, pages 1–8, June 2016.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural

networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014.

[25] J. Tang, J. Luo, T. Tjahjadi, and F. Guo. Robust arbitrary-view gait recognition based on 3d partial similarity matching. *IEEE Transactions on Image Processing*, 26(1):7–22, Jan 2017.

[26] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2016.

[27] C. Yan, B. Zhang, and F. Coenen. Multi-attributes gait identification by convolutional neural networks. In *2015 8th International Congress on Image and Signal Processing (CISP)*, pages 642–647, Oct 2015.

[28] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proceedings of 18th International Conference on Pattern Recognition (ICPR)*, volume 4, pages 441–444, 2006.

[29] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[30] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan. Robust view transformation model for gait recognition. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pages 2073–2076, Sept 2011.