Proceedings of the Fifth IIEEJ International Workshop
on Image Electronics and Visual Computing 2017
Da Nang, Vietnam, February 28 - March 3, 2017

MULTI-VIEW GAIT RECOGNITION BASED ON CONVOLUTIONAL NEURAL NETWORK

TzeWei Yeoh, Hernan Aguirre and Kiyoshi Tanaka
Faculty of Engineering,
Shinshu University, Japan.
Email: {15st213d@ ahernan@, ktanaka@}shinshu-u.ac.jp

ABSTRACT

Gait recognition is recognizing human through the style of walking of an individual. However, the presence of different covariate factors such as clothes, walking speeds, carrying objects, footwear and view angles complicates the task of recognition algorithms. Amongst of covariate factors, variations in viewing angles will drastically affect the overall performance of gait recognition. In recent, there has been significant research on this problem. However, conventional state-of-the-art methods have mostly use hand-crafted features for representing the human gait. In this work, we explore and study the use of convolutional neural networks (CNN) for automatically learn gait features or representations directly from low-level input raw data (i.e. *Gait Energy Image (GEI)*). Performance evaluations on the challenging view-invariant gait recognition of CASIA-B gait database, the experiment results show that our method provides competitive results, outperforming various conventional state-of-the-art methods.

1. INTRODUCTION

Robust and reliable human identification for surveillance and security access control has become highly demand these days [1]. Recently, human identification based on gait (i.e., identifying individuals by the style they walk), has attained considerable attention due to its ability to use it in automated visual security and surveillance systems [2]. The applicability of gait as biometric to use for recognizing subject's identity at a distance while being unobtrusive and non-invasive [1, 3].

However, in the real world application, robust and reliable gait recognition has become more complicated due to the existence of various covariate factors such as clothing, carrying condition, viewing angles, walking surface, elapsed time, etc. Amongst of covariate factors, the change of view angles has become one of the key problems in many gait recognition tasks. Variations in viewing angles will drastically affect the performance of gait recognition [4]. This problem has recently gained considerable attention from all the gait researchers.

Over the last few years, several gait approaches [5, 6] based on view transformation have been proposed which have the ability to deal with large viewing angle changes and do not rely on camera calibration. Recently, Makihara et al. [5] established a Singular Value Decomposition (SVD) based view transformation model to transform the gait features in probe viewing angle to that in gallery viewing angles. In [6], Kusakunniran et al. established a View Transformation Model (VTM) from the different point of view by adopting Support Vector Regression (SVR) technique. However, previous conventional approaches have mostly used hand-crafted features for representing human gait. This hand-crafted features to try to capture the essence of different visual of gait patterns. Although most of the conventional approaches have satisfactory performance, but the main drawback of these approaches are highly problem-dependent.

Recently, deep learning has gained significant attention from the computer vision community. This is because deep learning models are capable of learning multiple layers of feature hierarchies by constructing high-level features from low-level features [7]. Hence, they are more generic since the feature construction process is fully automated. Specifically, many recent studies have shown promising results for applying deep learning approaches to a variety of applications (e.g. image classification, text classification, natural language processing, scene labeling, etc.) [8–11].

The idea of exploring CNN features is also motivated by their applicable to different kinds of tasks. Therefore, in this work, we propose to use Convolutional Neural Networks (CNNs) to extract gait features and classify human through their gait at the same time.

We summarize our paper's contributions as follows.

- We present a CNN-based method for view-invariant gait recognition. The method will automatically learn to extract the most discriminative changes of gait features from a low-level input data (i.e. GEI) which is invariant to the variations in view angle change.
- We evaluate the performance of the proposed method

on the view-invariant of CASIA-B [4], and it achieves much higher performance than other conventional approaches.

The remainder of the paper is organized as follows. We will give a brief introduction to CNNs in Section 2, and then Section 3 presents the architecture of the proposed method. Experimental results and discussion are given in Section 4, and conclusion and future works of this paper is given

2. CNN OVERVIEW

The convolutional neural network (CNN or ConvNet) [12] model is a type of feed-forward artificial neural network which was inspired by biological processes [13] and are variations of multilayer perceptrons designed to use minimal amounts of preprocessing [14]. The connectivity pattern between its neurons was inspired by the organization of the animal cortex [15], whose individual neurons are tiled in such a way that they respond to overlapping regions in the visual field. A CNN architecture is usually formed by a sequence of layers that transform the input image volume into an output volume through a differentiable function. It is usually made up of a convolution layer, a spatial pooling layer, a normalization layer and followed by fully-connected layer. For a more detailed of the latest deep learning findings and the architecture of a CNN is designed, the interested readers are referred to [12, 16].

3. PROPOSED APPROACH

In this section we describe our proposed method to address the problem of gait recognition using CNN. The proposed framework for gait recognition based on CNN is represented in Fig. 2.

3.1. INPUT DATA

In this work, we use the Gait Energy Image (GEI) [17] as the gait feature descriptor and input data to the CNN. Example GEIs belonging to one subject are shown in Fig. 1. GEI is a spatiotemporal gait representation constructed using silhouettes. Given a size-normalized and horizontal aligned human walking binary silhouette sequenece $B(x, y, t)$, the grey-level GEI $G(x, y)$ is defined as follows

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B(x, y, t) \qquad (1)$$

where N is the number of frames in complete cycles of the sequence, x and y are values in the 2D image coordinate.

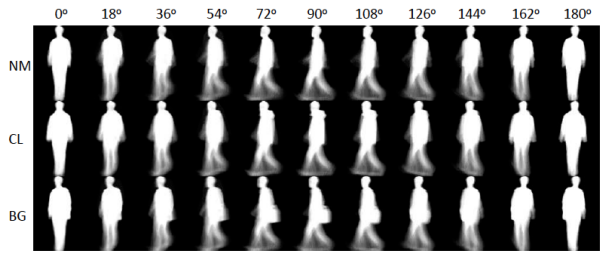

Figure 1: Examples of GEIs of the same subject in the CASIA-B dataset [4] 0°(left) to 180°(right) with an interval of 18°. Top row: normal walking (NM). Mid row: with a coat (CL). Bottom row: with a bag (BG).

3.2. CNN ARCHITECTURE

The overall architecture is shown in Fig. 2. The network contains 6 layers, with 3 convolutional layers (Conv1, 2 and 3), and 2 fully connected layer (FC4-5) and last follow by softmax layer. In our case, the first layer accepts 128x128x1 grayscale image which obtained from the sequence of GEI as input. The conv1, 2 and 3 yield 32, 64, 96 feature maps respectively. We set the number of neurons to 4096 and 2048 in FC4 and FC5, respectively. The input to the last layer, softmax layer, has n units, where n represents to the number of training subject samples taken from the dataset. Filter size for conv1 is set as 7x7 with strides of 1. While conv2 and conv3 use stride of 1, conv2 has filter size of 5x5, and conv3 has filter size of 3x3. Local Response Normalization (LRN) is employed after conv1, conv2 and conv3 with similar settings in [8]. Due to large number of parameters in the architecture, overfitting is inevitable. Hence, it is important to use dropout as form of regularizer. We follow the implementation shown in [8] to use the dropout after the FC4 and FC5 as most of the parameters are concentrated in these layers. We employed Rectified Linear Unit (ReLU) used as a activation function for all convolution layers, except for softmax layer that uses softmax regression as activation and act as a multi-class classifier for gait classification. The activation function has been proved to yield a better performance and speed up training time as reported in [8].

3.3. IMPLEMENTATION DETAILS

We trained our models using stochastic gradient descent (SGD) with mini-batch of size 40 and a momentum value of 0.8. The weights are initialized by using Gaussian distribution with zero mean and a standard deviation of 0.01 for all trainable layers. All the bias terms are initialized with the constant zero. We set our initial learning rate to 0.01. When a network has a too small learning rate will lead to slow convergence, while a too big learning rate makes the weights

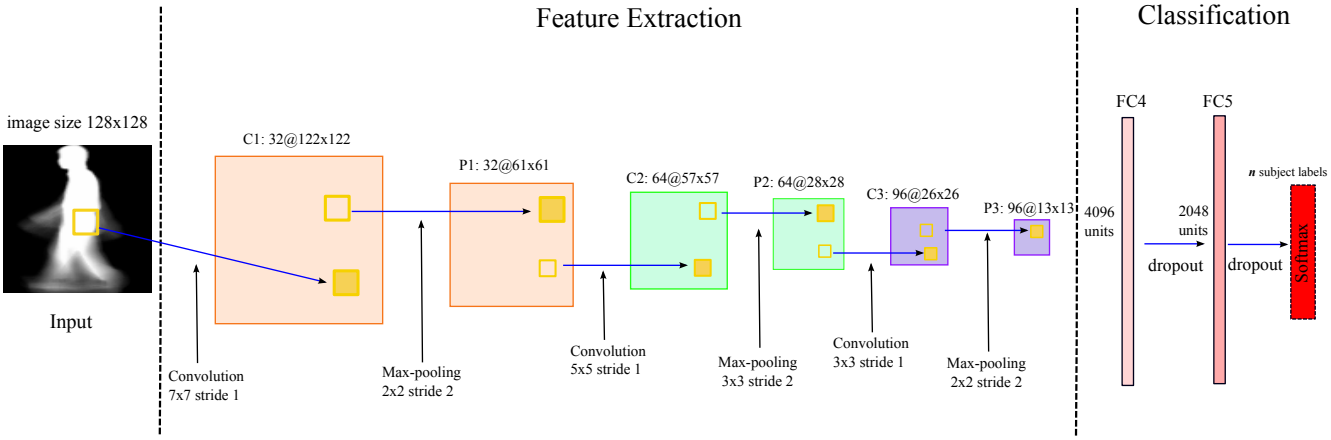

Figure 2: The architecture of our deep convolutional neural network for gait recognition.

and objective function diverge. We reduced the learning rate for all layers by factor of 10 every 1000 iterations prior to termination. We trained the model for 5000 iterations ($\sim 50 epochs$) for all experiments. We noticed that further training the model does not improve the results. The experiments was carried out using MatConvNet library [18] on NVIDIA GTX 970 4GB GPU. This library allows to prototype CNN architectures in an easy and fast manner using the Matlab environment. Besides, it takes advantage of CUDA and cuDNN [19] to improve the network's performance on classification tasks.

After we have acquired the gait features, the final stage (fully connected layer, FC5) consists of identifying those features to predict a subject identity. During this stage, a common set of similarity feature to individual subjects are computed using the Euclidean distance. Given a matching pair of a probe GEI P^i and gallery GEI G^i, the similarity score between them can computed through a trained network as follows:

$$d(P^i, G^i) = \sum_{N}^{n=1} \left| P^i(n) - G^i(n) \right| \qquad (2)$$

where $d(P^i, G^i)$ is a distance between the gait signatures P^i and G^i. N is size of gait feature vector. The smaller value of d means the higher possibility that the gait signature in between of the given matching pair, P^i and G^i are belong to the same person. Although the top layer of the CNN already comprises a softmax classifier, we can have the advantage of replacing the softmax layer with linear Support Vector Machine (SVM) classifier and using the extracted gait signatures as the input. Since our work is dealing with a multiclass problem, we use a binary SVM classifier with a linear kernel in a one-vs-all. Other papers (e.g. [20]) also indicated that this configuration of binary classifiers is suitable to obtain top-tier results in this problem. In our case, we L2-normalize the top fully connected layer before using it as a feature vector.

4. EXPERIMENTS AND RESULTS

To examine the effectiveness of our proposed method for view-invariant gait recognition, all our experiments are conducted on the largest multi-view gait CASIA-B [4] dataset.

4.1. DATASET DESCRIPTION

Experiments are conducted on CASIA-B gait dataset [4]. The dataset consists of the data from 124 subjects. The gait data was captured from 11 viewing angles, namely 0°, 18°, 36°, 54°, 72°, 90°, 108°, 126°, 144°, 162°, and 180°. There are 6 video sequences for each person under each different viewing angle. Therefore, we use a total of 8184 gait sequences.

4.2. RESULTS AND DISCUSSION

The performance of our proposed approach on CASIA-B dataset can be observed from Fig. 3. In overall, our method showed that these approaches can obtain good results when the data viewing angle is relatively small, i.e., 18°, however, the performance will eventually decreased due to view angle variations. From the experimental results, we can conclude the following key points. (1) CNN-SVM provides the highest accuracy than CNN-Softmax. (2) Being compared with conventional approach [6], the proposed method significantly improved multi-view gait recognition performance. The proposed method achieves the accuracy up to 90% for the close viewing angles (18°difference). However, the conventional approach [6] can only achieve up to 85% of accuracy.

Interestingly, CNN is able to identify human gait under the effects of view angle variations. In general, our experimental results using L2-SVMs show that by replacing

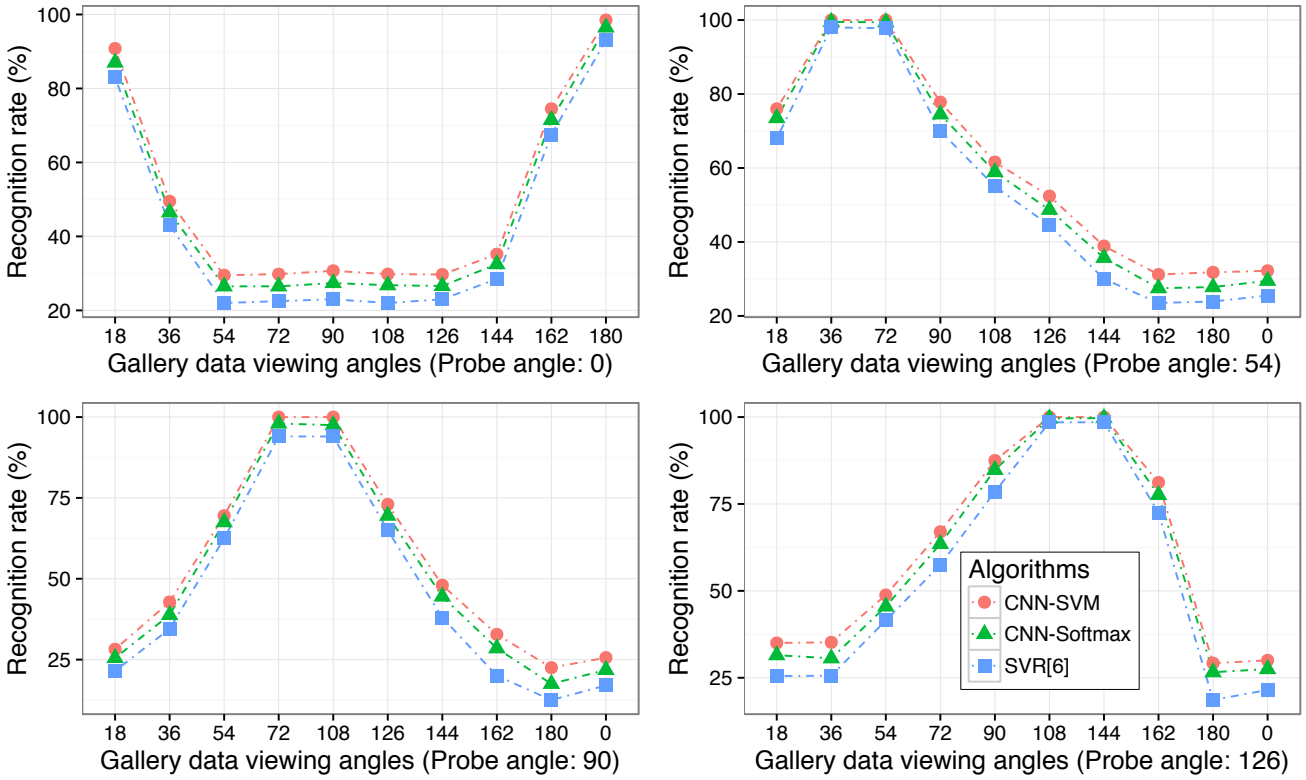

Figure 3: Performance evaluation on the CASIA-B dataset under the effects of view angle variation on gait identification.

softmax function with linear SVMs gives significant performance gains especially when under large viewing angles variation. The performance of CNN in identifying people based on their gait signatures shows that CNN is able to learn and differentiate these non-representational patterns. In order to better understand why CNN works in identifying human based on their gait, further investigation is needed and we leave it for future work.

5. CONCLUSION AND FUTURE WORK

This paper presents the investigation of view-invariant gait recognition using convolutional neural network. The experimental validation has been carried out on the large-scale gait CASIA-B dataset, by using gait energy image (GEI) as an input to a CNN. We designed a structure of CNN so that it can extract a view-invariant and discriminative features from the input GEI.

As future work, we are interested in visualizing the features in the CNN layers [21] for better understanding of how CNN works in learning patterns in gait under various clothing types effect. In addition, we plan to extend our study by using our proposed approach to other large-scale datasets for gait recognition. Our research can also be extended to evaluate pre-trained CNNs such as AlexNet, GoogLeNet, ResNet and etc. It would be ideal to experiment whether these architectures can yield better classification accuracy as compared to the proposed approach. We also plan to investigate further the use deep network such as ResNet [22] for our problem. This is because recently it has been reported that by training using deeper networks can obtain promising results.

ACKNOWLEDGMENT

The authors would like to thank CBSR for providing access to the CASIA-B gait database.

REFERENCES

[1] T. W. Yeoh, S. Zapotecas-Martínez, Y. Akimoto, H. E. Aguirre, and K. Tanaka, “Feature selection in gait classification using geometric pso assisted by svm,” in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2015, pp. 566–578.

[2] T. Yeoh, S. Zapotecas-Martínez, Y. Akimoto, H. Aguirre, and K. Tanaka, “Genetic algorithm as-

sisted by a svm for feature selection in gait classification," in *Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on*. IEEE, 2014, pp. 191–195.

[3] I. Bouchrika, J. N. Carter, and M. S. Nixon, "Towards automated visual surveillance using gait for identity recognition and tracking across multiple non-intersecting cameras," *Multimedia Tools and Applications*, vol. 75, no. 2, pp. 1201–1221, 2016.

[4] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4. IEEE, 2006, pp. 441–444.

[5] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *European Conference on Computer Vision*. Springer, 2006, pp. 151–163.

[6] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 974–981.

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding*. Springer, 2011, pp. 29–39.

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

[9] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in Neural Information Processing Systems*, 2015, pp. 649–657.

[10] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[12] Y. B. Ian Goodfellow and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[13] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, no. 5, pp. 555–559, 2003.

[14] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, 2015.

[15] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[16] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e2, 2014.

[17] J. Han and B. Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 316–322, 2006.

[18] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*. ACM, 2015, pp. 689–692.

[19] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cudnn: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.

[20] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.

[21] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.