Computers, Materials & Continua CMC, vol.60, no.3, pp.1041-1054, 2019

Attention-Aware Network with Latent Semantic Analysis for Clothing Invariant Gait Recognition

Hefei Ling[1], Jia Wu[1], Ping Li[1,*] and Jialie Shen[2]

Abstract: Gait recognition is a complicated task due to the existence of co-factors like carrying conditions, clothing, viewpoints, and surfaces which change the appearance of gait more or less. Among those co-factors, clothing analysis is the most challenging one in the area. Conventional methods which are proposed for clothing invariant gait recognition show the body parts and the underlying relationships from them are important for gait recognition. Fortunately, attention mechanism shows dramatic performance for highlighting discriminative regions. Meanwhile, latent semantic analysis is known for the ability of capturing latent semantic variables to represent the underlying attributes and capturing the relationships from the raw input. Thus, we propose a new CNN-based method which leverages advantage of the latent semantic analysis and attention mechanism. Based on discriminative features extracted using attention and the latent semantic analysis module respectively, multi-modal fusion method is proposed to fuse those features for its high fault tolerance in the decision level. Experiments on the most challenging clothing variation dataset: OU-ISIR TEADMILL dataset B show that our method outperforms other state-of-art gait approaches.

Keywords: Gait recognition, latent semantic analysis, attention mechanism, attention-aware neural network, clothing-invariant, feature fusion.

1 Introduction

In recent years, how to develop intelligent algorithm for modeling biometric traits plays more and more important roles in human identification. Most of the static traits such as fingerprint and iris have been used in reality. But these traits are limited by distance and the interaction with subjects [Bouchrika, Carter and Nixon (2016)]. Comparing with these biometric features, gait is an important coarse feature about motion so that gait recognition is robust to low resolution. It can be captured from long distance scenarios without the cooperation of subjects. And at the same time, the amount of cameras installed in public places is explosive increasing which make gait recognition possible for crime surveillance

[1] HuaZhong University of Science and Technology, Wuhan, 430074, China.
[2] Queen's University, Belfast, UK.
[*] Corresponding Author: Ping Li. Email: lpshome@hust.edu.cn.

and prevention.

However, there are still many challenges for applying gait recognition in the real life. Robust and discriminative features are important for the task of human identification because of the existence of covariates (e.g., carrying condition, camera viewpoint, clothing, the variation of walking speed, walking surface and so on). From most of appearance-based gait recognition methods [Wu, Huang, Wang et al. (2016)], the variation of clothing and carrying condition affects the performance of gait recognition drastically. These co-factors take the same problems to clothing invariant gait recognition, they change the appearance of subjects greatly. So, it becomes a hotspot for researchers.

In order to tackle the problem of the variation of appearance caused by clothing variation. There are a wide range of methods proposed in recent years (for recent review [Lee, Belkhatir and Sanei (2014)]), most of conventional approaches use hand-crafted features to represent the clothing-invariant human gait. For example, Shariful et al. [Shariful, Islam, Akter et al. (2014)] proposed a method called random window subspace (RWSM) to split raw input into small window chunks to get the gait segmentation and contribution of each body part for clothing-invariant gait recognition. Guan et al. [Guan, Li and Hu (2012)] proposed a random subspace method (RSM) based on computing a full hypothesis space, the method randomly chooses subspaces for classification. And Hossain et al. [Hossain, Makihara, Wang et al. (2010)] proposed a part-based gait identification in the light of substantial clothing variations, which exploits the discrimination capability as a matching weight for each part and controls the weights adaptively based on the distribution of distances between the probe and all the galleries. Rokanujjaman et al. [Rokanujjaman, Islam, Hossain et al. (2015)] proposed an effective parts definition approach based on the contribution of each row when it merges orderly from bottom to top. It shows that some rows have positive effects and some rows have negative effects for gait recognition. Based on the positive and negative bias, they defined three most effective body parts and two redundant body parts. Discarding two redundant parts and considering only three effective body parts improve the performance of gait recognition effectively. Actually, the pipeline of most of the conventional methods for clothing invariant gait recognition is always dividing the body into components firstly, and learns the weights of the features from different components. But the performance of these methods are unsatisfied because of the inevitable errors in extracting local features by traditional methods. While, they show the importance of local information and the relationship among them.

Besides those conventional approaches, the deep learning approach [Yeoh, Aguirre and Tanaka (2017)] automatically learns clothing-invariant gait features directly from raw data. Convolutional neural networks make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies), so they give great performance in object recognition and are applied in many fields. Zhou et al. [Zhou, Liang, Li et al. (2018)] use deep learning method in road traffic sign recognition. It is obvious that the CNN-based approaches outperform those conventional methods in many aspects. The CNN-based methods are easier to capture the features from raw input. At the

same time, from the aforementioned conventional methods, the latent attributes and local features from limbs are important in the field of clothing invariant gait recognition. To take advantage of the CNN-based methods and make use of the advantages from conventional methods, a more effective method based on convolutional neural network is urgent to proposed.

Attention network [Zhao, Wu, Feng et al. (2017)] and latent semantic features [Li and Guo (2014)] play important roles in the field of computer version. Attention network learns to pay more attention in important local parts of images. And latent semantic analysis (LSA) is known for the ability of capturing latent semantic features. Many recent studies show satisfying results than previous classification network [Krizhevsky, Sutskever and Hinton (2012)] by applying attention mechanism and LSA. They perform well in a variety of applications such as scene classification [Li and Guo (2014)], natural language processing [Fei, Cai-Hong, Wang et al. (2015)] and so on.

Inspired by the excellent performance of attention mechanism and latent semantic analysis, we employ latent semantic features to help analyze the contribution for different parts of images and get the latent relationships among features and classification results. And attention-aware network captures more discriminative features which highlight the important regions from subjects. In this paper, we combine the advantages of attention mechanism and LSA respectively, and design a new CNN-based method to address the problem of clothing invariant gait recognition.

We summarize the contribution of our work as following:

Firstly, we propose a specific CNN-based method for clothing-invariant gait recognition. The method automatically learns to combine features extracted from low-level input and latent semantic features from middle-level features which get a good representation for clothing invariant gait recognition.

Secondly, we evaluate our method on the most challenging clothing variant dataset: OU-ISIR Treadmill B dataset which includes the different clothing conditions, and it achieves better performance than other sate-of-art methods.

In the remainder, we detail our paper as following: related work about attention mechanism, latent semantic analysis and gait recognition are introduced in Section 2. After Section 2, how do CNNs, latent semantic analysis and attention combine and work are demonstrated in Section 3. Then experimental results are shown in Section 4. Finally, we give a conclusion in Section 5.

2 Related work

Approaches to gait recognition can be classified into two categories, one is model-based [Shariful, Islam, Akter et al. (2014); Guan, Li and Hu (2012); Shen, Pang, Tao et al. (2010)] and the other is model-free methods [Wu, Huang, Wang et al. (2016)]. Model-based methods are always conventional methods considered to be made up of statics from shape of human bodies and the components that can reflect the dynamic features of a cycle of

gait. It is majoring in modeling the structure of human body. The other method extracts gait feature from the raw input without considering the structure of subjects, it focus on the the shape of the silhouette rather than fitting it to a chosen model. Our method combines the structure of human body with model-free method so it can remedy the dependencies of model-free approaches on clothing variation by attention mechanism and latent semantic analysis.

Attention mechanism [Wang, Jiang, Qian et al. (2017)] is designed to highlight discriminative features for various kinds of tasks including images classification [Cao, Liu, Yang et al. (2016)], semantic segmentation [Chen, Yi, Jiang et al. (2016)], image question answering [Yang, He, Gao et al. (2016)], image captioning [Mnih, Heess, Graves et al. (2014)] and so on. Attention mechanism is effective in understanding images, since it adaptively focuses on related regions of the images when the deep networks are trained with spatially-related labels for capturing the underlying relations of labels and provides spatial regularization for the the results. In some extent, attention mechanism is similar to the conventional methods for clothing-invariant gait recognition but attention mechanism highlights the salient features automatically. Except the attention mechanism for gait recognition, there is a dramatic method to extract underlying attributes among those subjects. LSA learns latent features for gait recognition, which are important features and compensate the spatial features from attention-aware network.

LSA is a topic-model technique in neural language processing for improving information retrieval, it is first introduced by Deerwester et al. in 1988 [Deerwester (1988)] and further improved in 1990 [Deerwester (2010)]. Recently, the idea of latent semantic representation learning has been used in computer vision community. Zhiwu Lu proposed a novel latent semantic learning method for extracting high-level latent semantics from a large vocabulary of abundant mid-level features [Lu and Peng (2013)] for human action recognition. Bergamo et al. [Bergamo, Torresani and Fitzgibbon (2011)] applied a compact code learning method for object categorization, which uses a set of latent binary indicator variables as the intermediate representation of images. In the field of image retrieval and object detection, latent semantic learning can also be used to extract high-level features for latent semantic. It is obvious that features learned from latent semantic analysis extracting latent features not given before, and combining the features from improved CNN-based model with attention mechanism and latent semantic analysis can improve the performance of our task: clothing invariant gait recognition.

3 Methodology

We propose a convolutional neural network for clothing invariant gait recognition, which utilizes attention model for adaptive weights of different parts and latent semantic analysis for learning latent semantic features. The framework of our latent-attention compositional network (LACN) is illustrated in Fig. 1. The input data of our method is gait energy image (GEI) [Man and Bhanu (2005)], it is the average silhouette over one walking cycle of gait. And GEI is the most common input data for whether traditional methods or

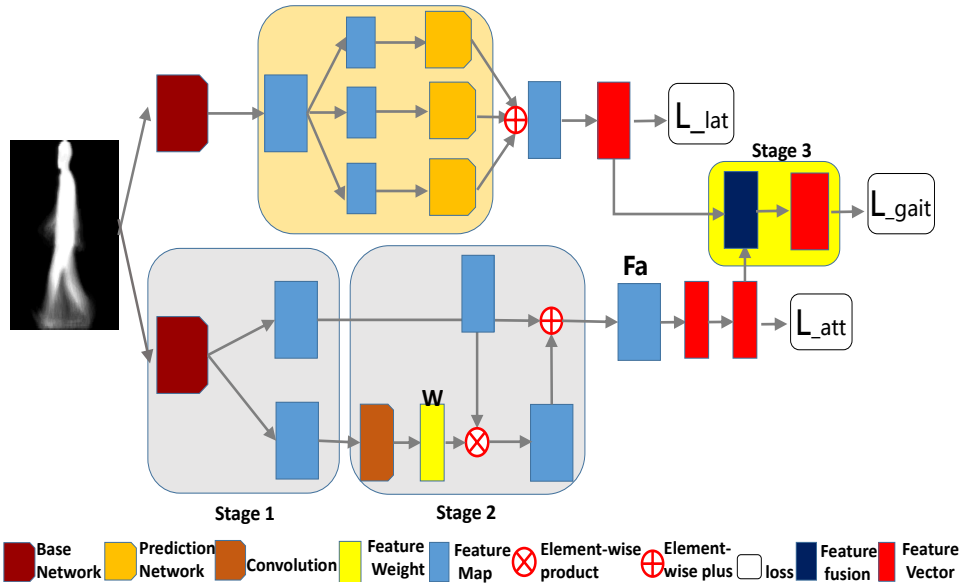

Figure 1: The pipeline of our network is illustrated in the figure. The base network is the same as the CNN-based method [Yeoh, Aguirre and Tanaka (2017)], which is composed of there convolutional layers, the kernel size are 7×7, 5×5 and 3×3 respectively. After capturing the feature maps, the attention module learns a soft mask and gets new features from the base network. In the latent semantic module, we divide the features from base network into fixed number of components and get latent variables for the corresponding components. Then, calculate the relationship with the final gait labels. Finally, we fuse the features from the two modules using convolutional layer with kernel size 1×1 to get discriminative and robust features

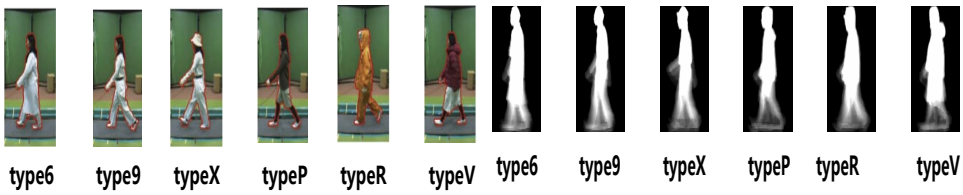

Figure 2: Samples of images from different kinds of clothing variations of OU-ISIR dataset B and the corresponding GEIs [Makihara, Mannami and Tsuji (2012)]

CNN-based methods. The samples and corresponding GEIs from dataset of different clothing combination are illustrated in Fig. 2. LACN consists of two main components: one combines the attention mechanism with latent semantic analysis for multi-level feature extracting, the other is multi-modal fusion which fuses the features from different feature extracting modules.

The attention model pays attention to high-level representation for the whole input data. It is constructed by two-branch convolutional neural network. Latent semantic analysis is used for extracting middle-level features that are ignored in high level. Finally, the features fusion strategy combines the features from different levels. The details for the two components are discussed in next three Subsections (3.1, 3.2 and 3.3).

Motivated by the conventional methods for clothing-invariant gait recognition. Dividing the input GEI into small fixed subspaces and getting latent variables from those subspaces is an effective way to get more discriminative features. As a result, we employ latent semantic analysis called patch-based latent semantic learning model for latent semantic features.

In this module, t labeled $\{X^i, Y^i\}_{i=1}^{n}$ images are given, where the X^i denotes the $i-th$ image and Y^i is the label for the image. We aim to learn a model from X^i to Y^i, the first step is to divide the input GEI into non-overlapping patches, the patches forms low-level features of input GEI, the features from these patches are regarded as latent variables Z_i^j. To predict the results from those latent variables, we take the each Z_i^j as latent high-level visual features, and get the gait label by the summarizing the high-level visual features inferred from their corresponding patches.

It is obvious that the latent variables are predicted from the input GEI. In theory, they can also represent the discriminative high-level features for the target gait labels. From the assumption, we formula the two stages of the prediction problems as the following unified optimization over the loss function.

3.1 Latent semantic analysis

$$min_{Z^i,W} \sum_{i,j} (L(y^i, f(\sum_j Z_i^j; W))) \tag{1}$$

where the function $f(*)$ is the function that predicts the gait labels from the latent variables

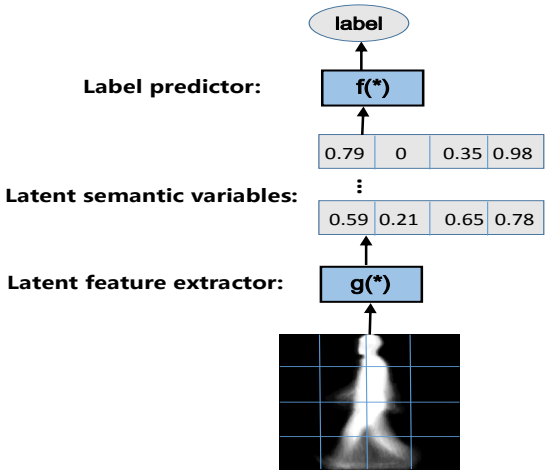

Figure 3: The procedure of the latent semantic analysis

from the whole image, W denotes the model parameters of the prediction function. Latent variable Z^{j} is computed by the latent features extractor, formulated as the Eq. (2).

$$Z_i^j = g(X_i^j;\theta)Z^j \in [0,1]^{n\times m} \quad for \quad j = 1,.....,t \tag{2}$$

The process to extract latent semantic features and capture the final result from those latent variables are demonstrated in Fig. 3, the procedure of function $f(*)$ and $g(*)$ are linear function as Eqs. (3) and (4) respectively.

$$g(X_i^j;\theta) = X_i^i\theta^{\mathrm{T}} + b^{\mathrm{T}} \tag{3}$$

$$f(\sum_j Z_i^j;W,b))) = \sum_j Z_i^j W + b^{\mathrm{T}} \tag{4}$$

From those fixed patches, latent variables are calculated to the corresponding patches and improve the performance of prediction function at the same time.

3.2 Attention model for adaptive weights of features

Attention maps highlight discriminative regions of different parts from human body. The attention network stimulates selection from feature maps by a soft mask which includes the weights of every dimension of features. As shown in Fig. 1, we design an attention-aware structure to capture specific regions from GEI. There are two chunks for the attention model. The one learns a soft mask for the feature maps from the base network which extracts features automatically by the other main chunk. The soft mask highlights the regions from corresponding part and plays a important role for its robust features.

Feature maps from the main chunk of input GEI are defined as Eq. (7).

$$X_{base} = f_{base}(I_{GEI}, \theta_{base}) \tag{5}$$

where I is the input data (GEI). To the result better than the original features X. Then, the second stage refines the attention maps A by modifying all previous prediction, θ_{att} is the parameters learnt from the attention modules. The attention module consists of two layers (the first layer has 512 filters with kernel size 1×1 and the second is sigmoid layer).

$$A = \frac{exp(f_{att}(X_{base_{i,j}}, \theta_{att}))}{\sum_{i,j} exp\{f_{att}(X_{i,j}, \theta_{att})\}} \tag{6}$$

The result from attention maps ranges from 0 to 1, it represents how important the original features is. The outputs F of final result are formula as,

$$F_{att} = (1 + A) \times X_{base} \tag{7}$$

From the formulation, it is obvious that attention map works as discriminative features selector which selects the original features X. Although attention maps adaptively capture the salient features. So the loss for attention modules is:

$$L_{att} = L^{att}(F_{att}) \tag{8}$$

where L^{att} denotes the loss function of confidence maps from attention-aware network, it is cross entropy loss.

We emphasize that the attention model calculates soft weights for feature maps from subjects, and it allows the gradient of loss function to be back-propagated through. The output A from attention module is actual a mask for the corresponding feature map F which adaptively highlights the important components of subjects. From Fig.4, the attention module highlights the limbs and head of subjects, which are discriminative parts in the problem of clothing-invariant gait recognition.

3.3 Feature fusion and classification

To fuse the feature from network with attention mechanism and latent semantic analysis and get better performance from the two modules, we joint the two kinds of features. Here we will introduce how we get the new features and calculate the final result from new features.

Features from attention-aware network f_{att} and latent semantic analysis f_{latent} are multi-modal features. After jointing the f_{att} and f_{latent} by channels, we can get the final features f_{fin}, and employ a convolutional layer with kernel size 1×1 to get higher-level features f_{mix} from the two kinds of features. After the feature extracting, we use the features f_{mix} to calculate the similarity of individual subjects using the Euclidean distance.

$$d(P^i, G^i) = \sum_{N}^{n=1} ||P^i(n) - G^i(n)|| \tag{9}$$

where $d(P^i, G^i)$ is a distance between the images from gallery and probe, N is the size of feature vectors. The smaller the value of d the higher possibility of the given matching pair and find the corresponding subject with the highest similarity in the gallery.

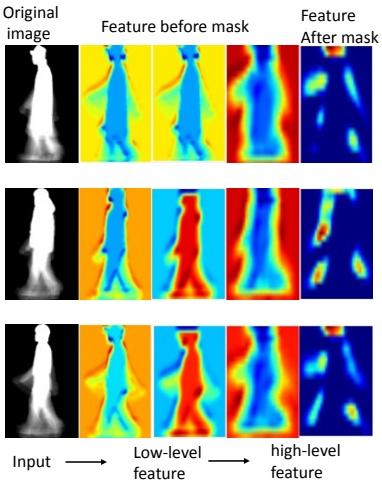

Figure 4: Example images illustrate that different features have different corresponding attention masks in our network. As we can seen in the figure, the attention chunk highlights the limbs and head of human body which are robust from the changing of appearance caused by clothing variation

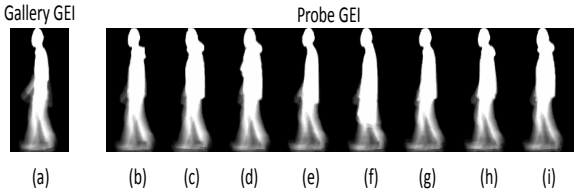

Figure 5: Samples of evaluation set, the image in the left (a) is in normal clothing type is used as gallery images, images in the right (b)-(i) are probe set with different clothing combinations

4 Experiments

4.1 Database description

The proposed method is evaluated on OU-ISIR Treadmill dataset B [Makihara, Mannami and Tsuji (2012)]. OU-ISIR Treadmill dataset B is a large gait dataset for evaluation of gait methods in presence of variations in clothing. It includes 68 subjects with up to 32 types of clothing combinations. Tab. 1 shows 15 different types of clothes used in constructing the dataset. Tab. 2 shows clothing combinations based on the 15 different types of clothes. For the most common approaches, the setup for the dataset is split into three parts including training set, probe set and gallery set. And there are 446 samples of 20 subjects from all types of clothing combination in training set, 48 sequences of 48 different subjects from normal clothing type, the probe set is consist of the rest clothing types of the 48 subjects

Table 1: List of clothes used in OU-ISIR treadmill dataset B [Makihara, Mannami and Tsuji (2012)]

RP-Regular pants	BP-Baggy pants	CW-Casual wear
Sk-Skirt	CP-Casual pants	HS-Half shirt
LC-Long coat	Pk-Parker	DJ-Down jacket
SP-Short pants	Ht-Hat	FS-Full shirt
Cs-Casquette cap	RC-Rain coat	Mf-Muffler

except samples in gallery set. The total number of rest clothing types of these 48 subjects is 856. But this kind of setup is not suitable for deep learning approaches, one reason is that the clothing type in training set in not complete from all kinds of clothing types, the other is that 446 sequences for training set is not enough for the input of deep learning approaches. To capture the discriminative feature from the variant clothing types, 32 kinds of clothing types and enough data for input of deep learning are necessary for training. So, in our work, the whole dataset are divided into two parts, the one is used to train the model the other is for evaluation. And the proportion for training and evaluation is 80/20 respectively. The subjects from the two subsets are not overlapping, and sequences in normal clothing type from all subjects in the evaluation are used for gallery set, probe set are composed of the rest data from evaluation. The samples from gallery and probe set are illustrated in Fig. 5.

Table 2: Different clothing combinations used in the OU-ISIR B dataset [Makihara, Mannami and Tsuji (2012)] (Abbreviation: Clothes type ID)

type	s1	s2	s3	type	s1	s2	type	s1	s2
3	RP	HS	Ht	0	CP	CW	F	CP	FS
4	RP	HS	Cs	2	RP	HS	G	CP	Pk
6	RP	LC	Mf	5	RP	LC	H	CP	DJ
7	RP	LC	Ht	9	RP	FS	I	BP	HS
8	RP	LC	Cs	A	RP	Pk	J	BP	LC
C	RP	DJ	Mf	B	RP	Dj	K	BP	FS
X	RP	FS	Ht	D	CP	HS	L	BP	Pk
Y	RP	FS	Cs	E	CP	LC	M	BP	DJ
N	SP	HS	-	P	SP	Pk	R	RC	-
S	Sk	HS	-	T	Sk	FS	U	Sk	PK
V	Sk	DJ	-	Z	SP	FS	-	-	-

4.2 Performance evaluation

1) Performance analysis with clothing variations effect

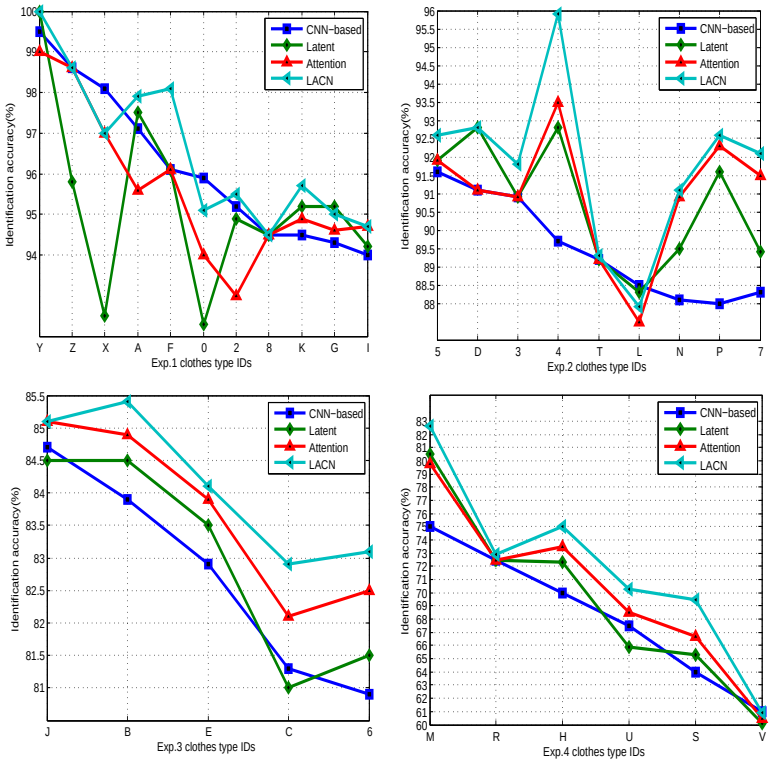

Figure 6: Performance of our method and state-of-art CNN-based methods on OU-ISIR Treadmill B dataset under the 32 different clothing combination

To demonstrate the effectiveness for our method, we conduct experiments on the dataset: OU-ISIR Treadmill B. The results of two kinds of features extracting from two modules and the final features are illustrated in Fig. 6. From the results, we can observe the experiments' results, we can observe that there are four level difficulties of clothing combination in the dataset OU-ISIR Treadmill B. In the experiment 1-4 (Exp.1-4), the CNN-based [Yeoh, Aguirre and Tanaka (2017)] method is the base network of our proposed method. The performance of attention module and latent semantic analysis module are better than CNN-based method in most of clothing types. What is more, our proposed method which combines the two modules outperforms the two modules respectively and it also shows better results than CNN-based method especially in the clothes type 4 (regular pants and half shirt) and M (baggy pants). It proves that the two-level features compensate for each other.

2) Comparison with state-of-art methods

In the experiment, we evaluate our method on the test set of dataset, and calculate the average accuracy. Compared our method with some state-of-art methods, Tab. 3 summarize the comparison of results with the hand-craft methods [Shariful, Islam, Akter et al. (2014);

Guan, Li and Hu (2012)], CNN-based method [Yeoh, Aguirre and Tanaka (2017)] and our method. It shows our method achieve better performance than state-of-art methods.

Table 3: List of clothes used in OU-ISIR treadmill dataset B [Makihara, Mannami and Tsuji (2012)]

Methods	Accuracy
RWSW [Shariful, Islam, Akter et al. (2014)]	78.54
RSW [Guan, Li and Hu (2012)]	80.44
CNN-based [Yeoh, Aguirre and Tanaka (2017)]	87.8
Our method	89.2

5 Conclusion

In this paper, we combine latent semantic analysis and attention mechanism for clothing-invariant gait recognition to get robust and discriminative features end-to-end. And fuse them for higher-level representation which improves the performance of gait recognition. The proposed method not only makes use of the advantages of CNN-based method which learns high-level feature from raw input data but also highlights the important regions from subjects. Local information is emphasized by attention mechanism in our method. At the same time, latent semantic variables play an essential role in our method, the number of latent variables are not the more the better, here we chose 30 variables after comparing the performance of the gait recognition. The performance of our method also shows it outperforms the state-of-art methods.

In our future work, we take additive sequential information into consideration. Although GEI is most popular representation for gait, but it obviously loses spatial and sequential information in some extent. To make use of sequential information, the raw input can be a cycle of silhouette or raw images. So the network for extracting sequential information is suitable for clothing-invariant. Attention-based long short term memory network (LSTM) [Greff, Srivastava, Koutnik et al. (2017)]) is the next step of our future work.

Acknowledgement: This work was supported in part by the Natural Science Foundation of China under Grant U1536203, in part by the National key research and development program of China (2016QY01W0200), in part by the Major Scientific and Technological Project of Hubei Province (2018AAA068).

References

Bergamo, A.; Torresani, L.; Fitzgibbon, A. W. (2011): Picodes: learning a compact code for novel-category recognition. *Advances in Neural Information Processing Systems 24*, pp. 2088-2096.

Bouchrika, I.; Carter, J. N.; Nixon, M. S. (2016): 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.

Cao, C.; Liu, X.; Yang, Y.; Yu, Y.; Wang, J. et al. (2016): Look and think twice: capturing top-down visual attention with feedback convolutional neural networks. *IEEE International Conference on Computer Vision*, pp. 2956-2964.

Chen, L. C.; Yi, Y.; Jiang, W.; Wei, X.; Yuille, A. L. (2016): Attention to scale: scale-aware semantic image segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3640-3649.

Deerwester, S. (1988): Improving information retrieval with latent semantic indexing. *Information Sciences*, vol. 100, no. 1-4, pp. 105-137.

Deerwester, S. (2010): Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407.

Fei, W.; Cai-Hong, L. I.; Wang, J. S.; Jiao, X. U.; Lian, L. I. (2015): A two-stage feature selection method for text categorization by using category correlation degree and latent semantic indexing. *Journal of Shanghai Jiaotong University (Science)*, vol. 20, no. 1, pp. 44-50.

Greff, K.; Srivastava, R. K.; Koutnik, J.; Steunebrink, B. R.; Schmidhuber, J. (2017): Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222-2232.

Guan, Y.; Li, C. T.; Hu, Y. (2012): Robust clothing-invariant gait recognition. *Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 321-324.

Hossain, M. A.; Makihara, Y.; Wang, J.; Yagi, Y. (2010): Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, vol. 43, no. 6, pp. 2281-2291.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012): Imagenet classification with deep convolutional neural networks. *International Conference on Neural Information Processing Systems*, pp. 1097-1105.

Lee, T. K.; Belkhatir, M.; Sanei, S. (2014): A comprehensive review of past and present vision-based techniques for gait recognition. *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2833-2869.

Li, X.; Guo, Y. (2014): Latent semantic representation learning for scene classification. *International Conference on International Conference on Machine Learning*, pp. 532-540.

Lu, Z.; Peng, Y. (2013): Latent semantic learning with structured sparse representation for human action recognition. *Pattern Recognition*, vol. 46, no. 7, pp. 1799-1809.

Makihara, Y.; Mannami, H.; Tsuji, A. (2012): The ou-isir gait database comprising the treadmill dataset. *IPSJ Transactions on Computer Vision and Applications*, vol. 4, pp. 53-62.

Man, J.; Bhanu, B. (2005): Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316-322.

Mnih, V.; Heess, N.; Graves, A.; kavukcuoglu, k. (2014): Recurrent models of visual attention. *Advances in Neural Information Processing Systems 27*, pp. 2204-2212.

Rokanujjaman, M.; Islam, M. S.; Hossain, M. A.; Islam, M. R.; Makihara, Y. et al. (2015): Effective part-based gait identification using frequency-domain gait entropy features. *Multimedia Tools and Applications*, vol. 74, no. 9, pp. 3099-3120.

Shariful, M.; Islam, M. R.; Akter, M. S.; Hossain, M. A.; Molla, M. K. I. (2014): Window based clothing invariant gait recognition. *International Conference on Advances in Electrical Engineering*, pp. 411-414.

Shen, J.; Pang, H.; Tao, D.; Li, X. (2010): Dual phase learning for large scale video gait recognition. *Proceedings of the 16th International Multimedia Modeling Conference MMM*, pp. 500-510.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C. et al. (2017): Residual attention network for image classification. *Computer Vision & Pattern Recognition*, pp. 6450-6458.

Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; Tan, T. (2016): A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 209-226.

Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. (2016): Stacked attention networks for image question answering. *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 21-29.

Yeoh, T. W.; Aguirre, H. E.; Tanaka, K. (2017): Clothing-invariant gait recognition using convolutional neural network. *International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 1-5.

Zhao, B.; Wu, X.; Feng, J.; Peng, Q.; Yan, S. (2017): Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245-1256.

Zhou, S.; Liang, W.; Li, J.; Kim, J. U. (2018): Improved vgg model for road traffic sign recognition. *Computers, Materials & Continua*, vol. 57, no. 1, pp. 11-24.