Deep Spatial-Temporal Fusion Network for Video-Based Person Re-Identification

Lin Chen[1,2], Hua Yang[1,2], Ji Zhu[1,2], Qin Zhou[1,2], Shuang Wu[1,2], and Zhiyong Gao[1,2]
[1]Department of Electronic Engineering, Shanghai Jiao Tong University
[2]Institution of Image Communication and Network Engineering, Shanghai Jiao Tong University
(SJChenLin, hyang, zhiyong.gao, Zhou.qin.190, shuangwu)@sjtu.edu.cn, jizhu1023@gmail.com

Abstract

In this paper, we propose a novel deep end-to-end network to automatically learn the spatial-temporal fusion features for video-based person re-identification. Specifically, the proposed network consists of CNN and RNN to jointly learn both the spatial and the temporal features of input image sequences. The network is optimized by utilizing the siamese and softmax losses simultaneously to pull the instances of the same person closer and push the instances of different persons apart. Our network is trained on full-body and part-body image sequences respectively to learn complementary representations from holistic and local perspectives. By combining them together, we obtain more discriminative features that are beneficial to person re-identification. Experiments conducted on the PRID-2011, i-LIDS-VIS and MARS datasets show that the proposed method performs favorably against existing approaches.

1. Introduction

Person re-identification is of important capability for surveillance systems as well as human-computer interaction systems [1]. Yet it remains a challenging issue due to large variations in viewpoint and lighting across different views, as shown in Fig.1. Existing approaches mostly can be categorized into three types, single-shot, multi-shot, and video-based. Many traditional handcrafted feature extraction methods have been proposed in single shot case, by extracting low-level spatial appearance features such as colour, texture and intensity gradient histograms [26, 7, 25]. However, as can be seen in Fig.1 (a), single-shot appearance features were intrinsically limited due to the appearance changes from cross-view illumination variation, viewpoint difference, scale variation and background clutters. As for the multi-shot case, random shots from multiple views were chosen to detect appearance features and then utilized to do an average or linear superposition for a certain person,

Figure 1. Challenges and limitations on the different methods of person re-identification. (a) Single shot method, limited due to the appearance changes from cross-view illumination variation, viewpoint difference, scale variation and background clutters. (b) Multi-shot method, feature extracted from discontinuous person images contains no temporal information and with insufficient discrimination. (c) Video-based method reduces the influence of some ambiguous cases like occlusions with more continuous images to learn multiple visual features, and (d) extracts clear and intact temporal information which is more useful when existing cross-view illumination variations.

which provided more samples with different pose, viewpoint, and background for modeling human appearance, thus allowing a better model of the persons appearance to be built. But the feature extracted from discontinuous person images made the temporal information useless as well [8]. However, in real surveillance network, the visual appearance usually has a large variation due to clothes changing of the person images cross long period or different location, which makes the effect of recognizing a person only from spatial information to be not reliable enough.

Compared to methods on the still images (single, multi-shot method), video-based methods for the re-identification problem have more advantages in that they provide more images of the same person, which is useful for reducing the influence of some ambiguous cases and supplying more continuous appearance information. Existing handcrafted video-based methods usually lead to high-dimensional features. Some other algorithms investigated combinations of multiple visual features or space-time feature extraction, fusing them to get a better performance [20]. However, when existing cross-view illumination variations, the multiple visual features is not reliable enough than the temporal information of person. Introducing the temporal information can help to improve the re-identification performance since it contains more appearance evolution information. Representative works include gait recognition [11, 12, 15, 16] and temporal sequence matching [9, 18, 20]. But the temporal information is difficult to extract due to insufficient amount of data and lack of effective technology, which restrict the development of video-based method for person re-identification.

Recently, CNN (Convolutional Neural Networks) has been widely utilized to extract spatial information of the single-shot pedestrians and achieved success [22, 27, 17, 19, 1], since it establish multidimensional model of pedestrians and can provide a better and robust feature representation [4]. However, these method didn't consider the periodic features since CNN has no capacity of extracting correlation between the image sequence. On the other hand, the RNN (Recurrent Neural Networks) can learn long-term dependencies and remember information for long periods of time, thus it can accumulate the visual information as well as the evolution pattern of human appearance within the video sequence, yielding a sequence-level human feature representation [23]. However, RNN can't completely integrate all sequence frames' periodic information, especially earlier image frames. The output of RNN will easily lost some important information from the earlier person image frames. What's more, RNN has an aptitude only for comprehensive explicit periodicity, in that the temporal feature extracted from the whole person image sequences will easily lost detail of the local information such as gait information. Recently, siamese loss [3] and triplet loss [2] layer were proposed to be embedded into the deep learning architecture to reduce the intra-class variation and increase the inter-class variation, but these loss layer not fully took the label information of the data into account. McLaughlin et al. [14] and Wu et al. [21] proposed a deep recurrent network combining the RNN and CNN for the video-based person re-identification and achieved considerable performance. But they all only utilized the output of RNN as feature representation and not took the information loss by the RNN into account. And Wu et al. [21] utilized traditional metric learning method to improve the performance, demonstrating that the feature representation extracted from RNN left to be improved by the means of more effective method.

Inspired by the above mentioned works, we propose a novel end-to-end network architecture simply called CRF (CNN and RNN Fusion) specially for the video-based person re-identification. In our network, raw image sequences and optical flow information is used as input, CNN and RNN are subsequently combined into a unified architecture to effectively exploit and fuse both the spatial and temporal visual information, and a siamese incorporated softmax loss layer is built to fully utilized label information, pull positive pairs together and push negative pairs apart as well. Then, we train our model on full-body and part-body image sequences respectively to learn complementary holistic features and local region based features and finally obtain a more discriminative feature fusion representation. Our main contributions are summarized as twofold: 1) a novel end-to-end deep network for the video-based person re-identification, incorporating CNN, RNN and multi-loss layer (siamese and softmax loss) to jointly learn and fuse both the spatial and temporal information for a certain person; 2) an effective feature fusion method to simultaneously detect the spatial and temporal feature, fuse complementary holistic features and local region based features then finally obtain more discriminative features that are beneficial to person re-identification.

2. Our Method

The core idea of our CRF network is to jointly learn and fuse both the spatial and temporal feature of the person, and propose effective feature fusion method to obtain a more discriminative feature representation. In this way, we incorporate the CNN and RNN to respectively learn the spatial and temporal information, and train them on siamese and softmax loss layer to pull the instances of the same person closer and instances belonging to different persons farther. Then the model is trained and deployed on the full, upper and lower part of person images respectively to fuse complementary holistic features and local region based features together, learning the final discriminative feature representation. The diagram of our proposed CRF neural network architecture is shown in Fig.2. In following sections, we present the proposed CRF network and our feature fusion method in details.

2.1. Input

The input of the CRF network is same with the baseline proposed by the work [14], including raw images and the optical flow information. Since raw images encodes details of a persons appearance and clothing, adding optical flow

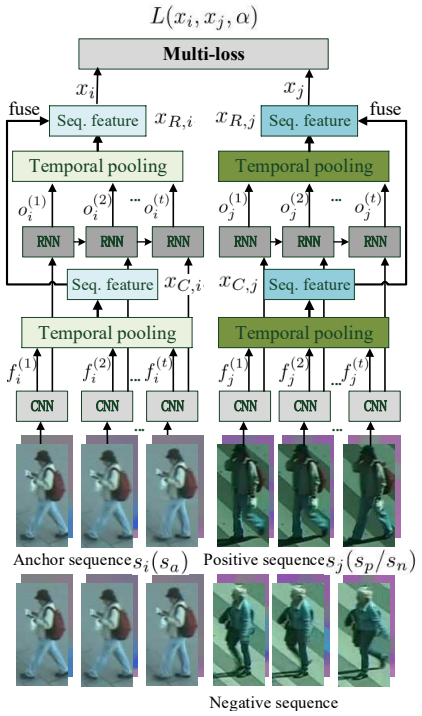


Figure 2. The illustration of our CRF network. Spatial features of image sequences detected after the CNN layer will be sent to a RNN layer to obtain the temporal information, then doing feature fusion, sending the final feature to the multi-loss (siamese loss and softmax loss) layer for optimization.

information will makes the details of a persons gait and other motion cues more clear [14].

2.2. CNN Layer

As illustrated in Fig.2, our network is first processed by a CNN layer to produce higher-level representation before the RNN layer, due to hierarchical feature extraction properties of deep networks layers [14]. After CNN layer, the feature vector output encodes details of a person's appearance and clothing, eliminating some interference and noise. Since RNN can accumulate the visual information and the evolution pattern of human appearance within the video sequence, we utilize better person appearance feature representation, i.e., the output of CNN, sending it into RNN layer to improve the performance of final temporal feature extracted by RNN. And in our case, we adopt three repetitive convolution layers with 5x5 kernel size, maxpooling layers and rectified linear unit (ReLu) activation layers. Given an input sequence $s = \left(s^0, \ldots, s^{T-1}\right)$, T is the sequence length, which is 16 in our case to better balance the number and length of the sequences, then s^t denotes the person image at time t. The operation of each CNN layer can be represented as $C^{'}(s^{(t)}) = ReLU(Maxpool(Conv(s^{(t)})))$.

The final feature vectors obtained after the CNN is

$$f^{(t)} = C(s^t), t \in 1 \cdots T. \tag{1}$$

where the C function refers to the simplification of feature extraction by the CNN layer.

2.3. Temporal pooling

The temporal pooling was proposed by [14] to aggregate information across all time steps and avoid bias towards later time-steps. In our case, as illustrated in Fig.2, the feature vectors $f^{(t)}$ from the CNN will then connect a temporal pooling layer to produce a single feature vector representing the appearance averaged over the whole input sequence of the person, i.e. spatial feature in our case, which can be denoted by following formula:

$$x_C = \frac{1}{T}\sum_{t=1}^{T} f^{(T)}. \tag{2}$$

2.4. RNN Layer

The output feature vector from the RNN layer then can synthesize anterior images within the video sequence. Specifically, given an input sequence $s = \left(s^0, \ldots, s^{T-1}\right)$, for s^t denotes the image in time t, the input of the RNN is $f^{(t)}$ which is final feature vectors obtained after the CNN. Then the RNN will learn long-term dependencies and remember information for long periods of time on the following operations:

$$o^{(t)} = W_i f^{(t)} + W_s r^{(t-1)}, \tag{3}$$

$$r^{(t)} = ReLU(o^{(t)}), \tag{4}$$

where $r^{(t)}$ will remember information at the previous time-step and allowing information to flow time-steps, the $o^{(t)}$ will produces an output based on both the current input and information from the previous time-steps as well. In our case, the $o^{(t)}$ and $f^{(t)}$ all through a linear combination to produce a feature vector with a dimension N, where N is 128. Then the $o^{(t)}$ will connect a temporal pooling layer to produce a single feature vector accumulating the appearance information of sequences to gain the periodic characteristics of our person image sequence, i.e., temporal feature such as gait or other pattern of human appearance, which can be denoted as

$$x_R = \frac{1}{T}\sum_{t=1}^{T} o^{(T)}. \tag{5}$$

2.5. Spatial-Temporal Feature Fusion

Although RNN network can remember information for long periods of time, it can't completely integrate all sequence frames' periodic information, especially earlier image frames. [14, 21] adopt temporal pooling to do an average with the earlier frames' output of RNN, which can

not commendably solve this problem. In this paper, we proposed an solution to this problem. Since the input of the RNN is the feature extracted by CNN layer, adding the information extracted by the CNN can make up for the lost information to a certain extent. In our paper, just as shown in Fig. 2, the final feature representation x_F is the fusion of the spatial and temporal features, i.e. the CNN output $f^{(t)}$ and RNN output $o^{(t)}$, the fusion operation is simply calculated as

$$x_F = x_C + x_R. \tag{6}$$

Then we send the final feature representation for further optimization.

2.6. Multi-Loss Layer

In our CRF network, we adopt siamese loss incorporated with softmax loss layer for optimization, to fully utilized label information, pull positive pairs together and push negative pairs apart as well. The siamese loss layer for person re-identification is proposed by [3]. Their core concept is dividing the input images into pairs, label them with 1 or -1 to tell the network the image pair is positive or negative. In our case, the positive pair contains two sequences that from a same person under different camera while the negative pair contains two image sequences that come from different person under random camera. To balance the number of positive and negative sequence pairs, we set the equal number of positive and negative pairs in each iteration. Specifically, Given an input sequences pair (s_i, s_j), feature representation (x_i, x_j) is obtained after our CRF network, and the loss function is calculated in following formula:

$$L_{sia}(x_i, x_j) = \begin{cases} \frac{1}{2}\left\|x_i - x_j\right\|_2^2, & i = j \\ \frac{1}{2}\left[max(\alpha - \left\|x_i - x_j\right\|_2^2, 0)\right]^2, & i \neq j \end{cases} \tag{7}$$

$$L_{sof} = -Wx_{y_i} + log\sum_j e^{Wx_j}, \tag{8}$$

$$L = W_1 L_{sia}(x_i, x_j) + W_2 L_{sof}(x_i) + W_3 L_{sof}(x_j), \tag{9}$$

where $\|.\|_2^2$ is the Euclidean distance between the feature vectors. L_{sia} denotes siamese loss while L_{sof} denotes softmax loss, and the final loss is calculated in formula 9. In our case, the α of the siamese loss is 2.0 to balance decline rate of the loss and the final accuracy performance. The loss weight W_1,W_2 and W_3 all is 1 to ensure correct classification and maximizes the relative distance between feature expression of negative pairs, minimizes which between positive pairs.

2.7. Local-Global Feature Fusion

RNN has an aptitude for comprehensive explicit periodicity. However, the whole person sequence image contains interference and noise, because the trajectory of upper limb is easy to change than lower limb. Utilizing RNN to detect periodic features of the whole image will lose detail of the local information, and its trajectory detail is not so obvious such as gait information. Based on this observation, we divide the person image into the upper part and lower part, and train separate deep models for them. For the same person, we then do fusion on the feature from the upper and lower part to gain part-based features x_{upper}, x_{lower}, i.e. local feature. Then fusing it with the global feature x_{global} extracted on the full image sequences to obtain final feature representation x_{final}, the performance is further improved, as shown in Fig. 3 and following formula:

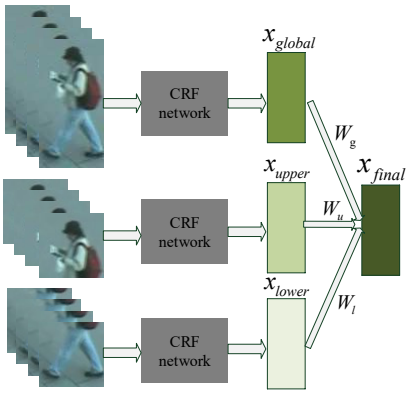


Figure 3. The illustration of our local and global feature fusion method. CRF denotes our CNN, RNN deep neural spatial-temporal feature fusion network. Training the same person with their whole and part images in the CRF network to detect their local and global features. Fusing them then gain the final better feature representation.

$$x_{final} = W_u x_{upper} + W_l x_{lower} + W_g x_{global}, \tag{10}$$

where the fusion weight W_u, W_l, W_g is adaptive parameters learned from the training sets.

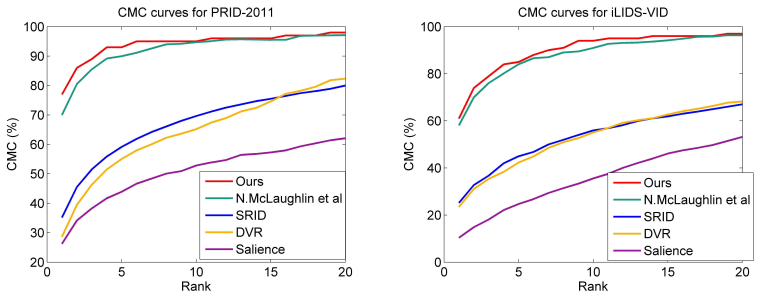
3. Experiments and Results

3.1. Evaluation Datasets

In this paper, we adopt three typical datasets widely used on the problem of video-based person re-identification to evaluate the performance of our method.

PRID-2011: The PRID 2011 [5] re-identification dataset contains two non-overlapping camera views, but only 200 people from the two camera views are adjacent. There are 400 image sequence pair on the dataset. Each image sequence has variable length from 5 to 675 and with an average number of 100. Compared with the iLIDS-VID dataset, it is captured in uncrowded outdoor scenes, more simple and clean background and rare cluttered occlusions.

i-LIDS-VID: The iLIDS-VID dataset [29] is captured at an airport arrival on two non-overlapping camera views.



(a) (b)

Figure 4. Comparing our method with other state of the art approaches and (a) illustrates the CMC rank curves on the PRID-2011 datasets, (b) illustrates which on the i-LIDS-VID datasets.

Table 1. Comparison with state-of-the-art methods on different datasets.(%)

datasets	*PRID-2011*				*i-LIDS-VID*				*MARS*			
CMC Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
DVR [20]	29	55	66	83	23	42	55	68	–	–	–	–
SRID [6]	35	59	70	80	25	45	56	66	–	–	–	–
K.Liu et al[10]	64	87	90	92	44	72	84	92	–	–	–	–
TDL [24]	57	80	88	94	46	77	90	96	–	–	–	–
RCN [21]	69	88	93	96	56	88	96	98	–	–	–	–
N.M et al[14]	70	90	95	97	58	84	91	96	–	–	–	–
Zheng et al [27]	77	**94**	–	**99**	53	81	–	95	65	82	–	89
Ours	**77**	93	95	98	**61**	**85**	**94**	**97**	**71**	**89**	**93**	**96**

Table 2. Comparison with [27].(%)

datasets	*PRID-2011*				*i-LIDS-VID*				*MARS*			
CMC Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Zheng et al [27](+ Eucl.)	59	94	96	99	41	70	79	85	59	77	–	87
Zheng et al [27](+ XQDA)	77	94	–	99	53	81	–	95	65	82	–	89
Ours(+ Eucl.)	77	93	95	98	61	85	94	97	71	89	93	96

Each camera view contains coincident 300 people and totally 600 image sequence pairs. Each image sequence has variable length from 23 to 192, and an average number of 73. Due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and occlusions, this dataset is much more challenging than the PRID-2011 dataset.

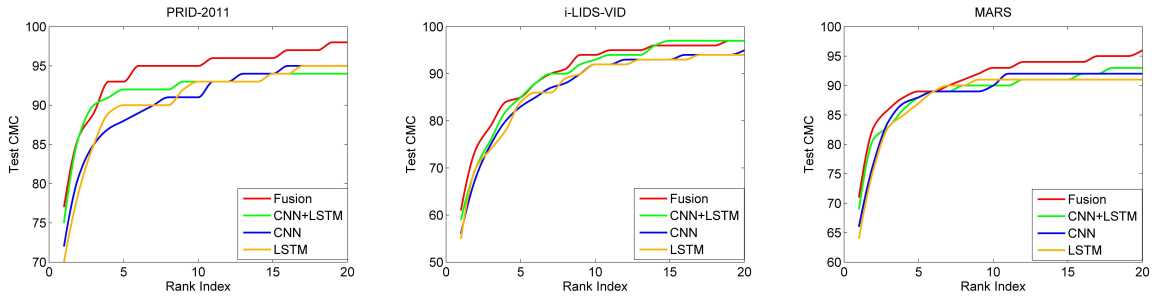
MARS: The MARS dataset [27] is the largest video re-identification dataset. As an extension of the Market-1501 dataset [28], it consists of 1261 different identities and 20715 tracklets. A large number of tracklets contain 25 to 50 frames, while most IDs are captured by 2 to 4 cameras and have 5 to 20 tracklets. Each identity has 13.2 tracklets on average. This dataset is much more large than the previous two datasets, remaining challenging and necessary to evaluate the performance of the proposed method.

3.2. Experimental Setup

Data preparation: Each dataset is randomly divided into two parts, one half for training and the other for testing. For example, there are 100 identities for training and the other 100 identities for testing on the PRID-2011 dataset. During training, all images are grouped into sequence pairs, i.e., positive and negative image sequence pairs with fixed length. To better balance the number and length of the sequences, we set the length as 16. In our experiment, we randomly choose a non-overlapping subset of 16 consecutive frames split from the full sequence to constitute each sequences. Therefore, each positive and negative sequence pairs are composed by 16 consecutive frames randomly selected from two full sequences of different cameras respectively.

To balance the number of positive and negative sequence pairs, data augmentation is applied to expand the datasets and increase the amount of negative pairs [13]. During testing, the same data augmentation is done on the probe and gallery data to ensure the consistency.

Training and testing: Each training procedure is more than 10 epochs, and more than 10 times to obtain an average and representative results. The dimension of the feature vector is 128 to decrease over-fitting. During testing, we choose all frames of a certain person and divide it into 16



(a) (b) (c)

Figure 5. Analysis the effectiveness of our proposed method. The CNN denotes only use feature extracted from CNN layer; RNN denotes only use features extracted after RNN layer; Then the CNN+RNN denotes the features extracted from the fusion layer after CNN and RNN; Fusion is the final feature integrated local feature (Upper, Lower) and global feature using our CRF network.

Table 3. Compare the performance of our different feature extraction methods on different datasets.(%)

datasets	*PRID-2011*				*i-LIDS-VID*				*MARS*			
CMC Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
CNN	72	88	91	95	56	83	92	95	66	88	90	92
RNN	70	90	93	95	55	84	90	94	64	87	91	91
CNN+RNN	75	92	93	94	59	85	93	97	69	88	90	93
Fusion	77	93	95	98	61	85	94	97	71	89	93	96

image frames one group to extract fusion features after CNN and RNN layer with pre-trained model, and extract final average feature to compute the matching rate against the features of gallery sequences under the simple Euclidean distance.

3.3. Evaluation and Results

Comparison with the state-of-the-art. We compare the performance of our proposed CRF network with several other state-of-the-art video-based methods, shown on Table 1 and Table 2, and the CMC curves is also shown in Fig. 4.

The baseline is the method proposed by the McLaughlin et al [14]. They proposed a DNN architecture but only detected the feature from the RNN layer. Compared with their method, in our method, information from CNN are subsequently combined to effectively make up for the lost information from RNN, while complementary holistic features and local based features are fused to obtain a more discriminative feature representation. From the results, our results all outperform other existing approaches, verifying that the features extracted from our method are favorably improved. On the more challenging i-LIDS-VID dataset, the improvement of our method compared with the baseline is less obvious than that in PRID-2011 and MARS. We think the reason is that the other two datasets is less challenging than iLIDS dataset and using only the feature from CNN or RNN can achieve a good performance, then our feature fusion strategy among these two kind of features shows more improvement. However, iLIDS-VID dataset is much more difficult which contains variations, cluttered background and occlusions. Only using the spatial or temporal feature does not work well for such challenges, which makes the performance of our fusion method limited as well. On the other hand, from the results on the work by [21] and [27], they even have some performance loss on the iLIDS-VID dataset compared with the work by [24] while better results on the PRID-2011. Our method all achieves better performance on these datasets no matter how challenging they are, which also demonstrates the effectiveness and robustness of our method. Compared with the state-of-the-art result proposed by [27], Our method focus on the feature extraction of ReID, not the metric learning, which is evaluated on Euclidean distance while the best results of [27] utilized metric learning like XQDA. As shown in Table 2, even with efficient metric learning method, there is no or limited superiority of [27] compared with our method. Other method such as RCN [21] not only utilized DNN model but also added traditional metric learning method (KISSME) to obtain good performance. Our method is evaluated on Euclidean distance and obtain better performance, which verifies that the features extracted from our method have obvious superiority and shows the effectiveness of our method as well.

Analysis of the proposed method. To better illustrate the effectiveness of our algorithm, we compare the performance of several baseline on these datasets, that is: extracting feature only by CNN layer, only by RNN layer, by CNN and RNN fusion layer, and by global-local feature fusion method based on our CRF network respectively. The comparison CMC result is shown on table 3 and Fig. 5.

RNN is known to have the problem of gradient vanishing, which makes it have better integration of adjacent

frames but can't completely integrate all sequence frames' periodic information, especially for earlier image frames. From the comparison of results, the performance of combining CNN and RNN is better than that only from CNN or RNN. The results confirms the effectiveness of our CRF network that adding the information extracted by the CNN to the RNN makes up for the lost information from the RNN layer. On the other hand, the feature from the CNN performs better than that of RNN due to the existing of the CNN and RNN fusion procedure, in which the CNN becomes dominated. Moreover, the performance of final global-local fusion feature is the best, verifying the complementarity between global features and local features and the effectiveness of our global-local feature fusion method. These results shown on the three datasets demonstrate the effectiveness of the temporal-spatial and local-global feature fusion method on our proposed CRF network.

4. Conclusion

In this paper, we propose a deep neural network architecture incorporating the convolution neural network (CNN) and recurrent neural network (RNN) to jointly exploit the spatial and temporal information for the video-based person re-identification. Then considering the complementarity between holistic features and local features, we fuse them together to finally obtain more discriminative features that are beneficial to person re-identification. The effectiveness of the our proposed method is verified on the three representative datasets, PRID-2011, i-LIDS-VID and MARS. Our method obtain favorable performance compared with the other existing state-of-the-art approaches. In future work, we plan to combine the deep learning with the traditional theory, making the deep learning more theoretical and interpretability.

5. Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC, Grant No. 61671289) and Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 15DZ1207403).

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

[2] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[3] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.

[4] A. Haque, A. Alahi, and L. Fei-Fei. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. *Person Re-identification by Descriptive and Discriminative Classification*. Springer Berlin Heidelberg, 2011.

[6] S. Karanam, Y. Li, and R. J. Radke. Sparse re-id: Block sparsity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2015.

[7] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013.

[8] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *British Machine Vision Conference*, 2015.

[9] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *2009 IEEE 12th international conference on computer vision*, pages 444–451. IEEE, 2009.

[10] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3810–3818, 2015.

[11] J. Man and B. Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2006.

[12] R. Martín-Félez and T. Xiang. Gait recognition by ranking. In *European Conference on Computer Vision*, pages 328–341. Springer, 2012.

[13] N. McLaughlin, J. M. Del Rincon, and P. Miller. Data-augmentation for reducing dataset bias in person re-identification. In *Advanced Video and Signal Based Surveillance, 2015 12th IEEE International Conference on*, pages 1–6. IEEE, 2015.

[14] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] M. S. Nixon, T. Tan, and R. Chellappa. *Human identification based on gait*, volume 4. Springer Science & Business Media, 2010.

[16] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, 27(2):162–177, 2005.

[17] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*, pages 732–748. Springer, 2016.

[18] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds

using dynamic time warping. In *European Conference on Computer Vision*, pages 423–432. Springer, 2012.

[19] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. *arXiv preprint arXiv:1605.03259*, 2016.

[20] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014.

[21] L. Wu, C. Shen, and A. v. d. Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*, 2016.

[22] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *arXiv preprint arXiv:1604.07528*, 2016.

[23] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016.

[24] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. *arXiv preprint arXiv:1604.08683*, 2016.

[25] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2528–2535, 2013.

[26] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[27] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[29] W. S. Zheng, S. Gong, and T. Xiang. Associating groups of people. *Active Range Imaging Dataset for Indoor Surveillance*, 2009.