See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-based Person Re-identification

Zhen Zhou[1,3] Yan Huang[1,3] Wei Wang[1,3] Liang Wang[1,2,3] Tieniu Tan[1,2,3]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
{zzhou, yhuang, wangwei, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

Surveillance cameras have been widely used in different scenes. Accordingly, a demanding need is to recognize a person under different cameras, which is called person re-identification. This topic has gained increasing interests in computer vision recently. However, less attention has been paid to video-based approaches, compared with image-based ones. Two steps are usually involved in previous approaches, namely feature learning and metric learning. But most of the existing approaches only focus on either feature learning or metric learning. Meanwhile, many of them do not take full use of the temporal and spatial information. In this paper, we concentrate on video-based person re-identification and build an end-to-end deep neural network architecture to jointly learn features and metrics. The proposed method can automatically pick out the most discriminative frames in a given video by a temporal attention model. Moreover, it integrates the surrounding information at each location by a spatial recurrent model when measuring the similarity with another pedestrian video. That is, our method handles spatial and temporal information simultaneously in a unified manner. The carefully designed experiments on three public datasets show the effectiveness of each component of the proposed deep network, performing better in comparison with the state-of-the-art methods.

1. Introduction

The person re-identification research aims to develop methods for matching pedestrian images/videos under two non-overlapping cameras. It draws increasing attention in computer vision because of a wide range of potential applications, such as the security in public places and criminal investigation. And more importantly, deep neural networks have shown to be effective for person re-identification, and have achieved much better performance than traditional methods [1, 6, 9, 12, 13, 18, 25, 36, 38].

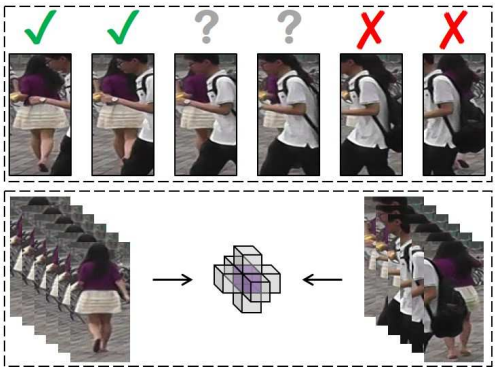

Figure 1: The *top* tries to explain that different frames in an image sequence provide different information. In this case, we want to recognize a girl wearing skirt. So "good" frames are those where the girl can be clearly watched, marked by green ticks. While "bad" ones are marked by red crosses and less important ones are marked by gray questions. The *bottom* is to illustrate that when matching two image sequences, we need to consider the surrounding pixels around each location. **The figures throughout this paper are best viewed in colors.**

In general, person re-identification methods can be classified into two categories, i.e., single static image-based approaches and the video-based ones. A large portion of existing works lies in the former category [7, 21, 28, 32, 33, 34] while only a few belong to the latter. Actually, video-based person re-identification is closer to practical applications since we do not need to manually pick out the desired im-

ages in a video, which are used to compare with images in another video. Furthermore, videos contain richer information than a single image [19, 20], which is beneficial for identifying a person under complex conditions, including occlusions, and the changes of illumination and viewpoint. Therefore, in this paper we will focus on video-based person re-identification.

Person re-identification methods generally involve two key steps, namely feature learning and metric learning. Feature learning aims at designing algorithms to generate discriminative features. Two pedestrian videos are matched if the distance/similarity between their features is the smallest/highest in the gallery set. While metric learning refers to develop metrics through which the similarity between two matched videos of the same pedestrian is higher than that between videos of different pedestrians. Most previous works on video-based person re-identification [22, 29, 37] pay attention to feature learning or metric learning independently. A recent trend is to design a deep neural network, say Convolutional Neural Network (CNN) [17] or Recurrent Neural Network (RNN) [26], to learn features [23, 39] or metrics [31]. In this paper, to leverage the merits of both feature learning and metric learning, we construct an end-to-end deep neural network architecture to learn them simultaneously.

As shown on the top of Figure 1, in a given image sequence, we observe that not all images are informative. The previous methods may be even confused if the occlusion is heavy. In this case, it is natural to expect that the desired person re-identification method can focus on those "good" images, which present relatively clear foreground. Hence in this paper we implement this idea by a temporal attention model (TAM), which exploits the temporal recurrent neural network [35] to assign a changeable weight to different frames in an image sequence. This enables the proposed method to selectively pay attention to more relevant images, thus further improving the performance of feature learning.

When comparing the similarity between two image sequences, the common way is to calculate the distance between their feature representations, which ignores the spatial difference during the sequence. As pictured on the bottom of Figure 1, in this paper the similarity between two corresponding location in a pair of image sequences is the integration of the surrounding information. Therefore, the proposed method is able to perform better metrics. We achieve this goal by the spatial recurrent model (SRM), which sweeps the image sequence along predefined directions.

We summarize the contributions of this work in three folds as follows.

1. Using the temporal attention model (TAM), we can measure the importance of each frame in a pedestrian video, which is useful for choosing more informative frames and thus improving feature learning.
2. The spatial recurrent model (SRM) is beneficial for exploring contextual information, which has been experimentally demonstrated effective for metric learning.
3. Feature learning and metric learning are incorporated into an end-to-end deep architecture, together with the aforementioned TAM and SRM, which achieves better results than the state-of-the-art methods.

The rest of this paper is organized as follows. In Section 2, we will review related works. Section 3 will first present the overall architecture of the proposed method, and then explain each important component in more details. Experimental results on three public datasets will be given in Section 4. At last we conclude this paper in Section 5.

2. Related Work

In this section, we first review some related works in person re-identification, especially those video-based methods and deep neural network based methods. Then we describe related works about spatial RNNs and temporal RNNs.

2.1. Person Re-identification Methods

Wang *et al*. [29] aim at selecting discriminative video fragments. They firstly choose the frames with the maximum or minimum flow energy, which is computed by optical flow fields. These selected frames, together with their contextual frames, comprise the so-called video fragments. HOG3D [16] is chosen as the feature extraction method for each video fragment. The similarity between two videos is the highest similarity between their video fragments. Liu *et al*. [22] try to extract features that encode the spatially and temporally aligned appearance of a pedestrian. They firstly detect the walking cycles by using the regulated flow energy profile and then split the whole video into segments. As to spatial alignment, the human body is described by six rectangles corresponding to the different body parts.

McLaughlin *et al*. [23] build a CNN to extract features of each frame and then apply an RNN to exploit the temporal information. A temporal pooling layer is adopted to summarize the output feature at each time step as the final representation. The input is a pair of image sequences and their corresponding optical flows. Besides the matching loss, each stream has an individual identity loss. Wu *et al*. [31] present a similar network architecture. Given a pair of pedestrian videos, they jointly train the convolution network and the recurrent layer to learn spatial-temporal features and the corresponding similarity.

Wang *et al*. [28] evaluate two different strategies for image-based person re-identification, i.e., single-image and cross-image representations. The former is learned by an identity classification task and the latter is obtained by a

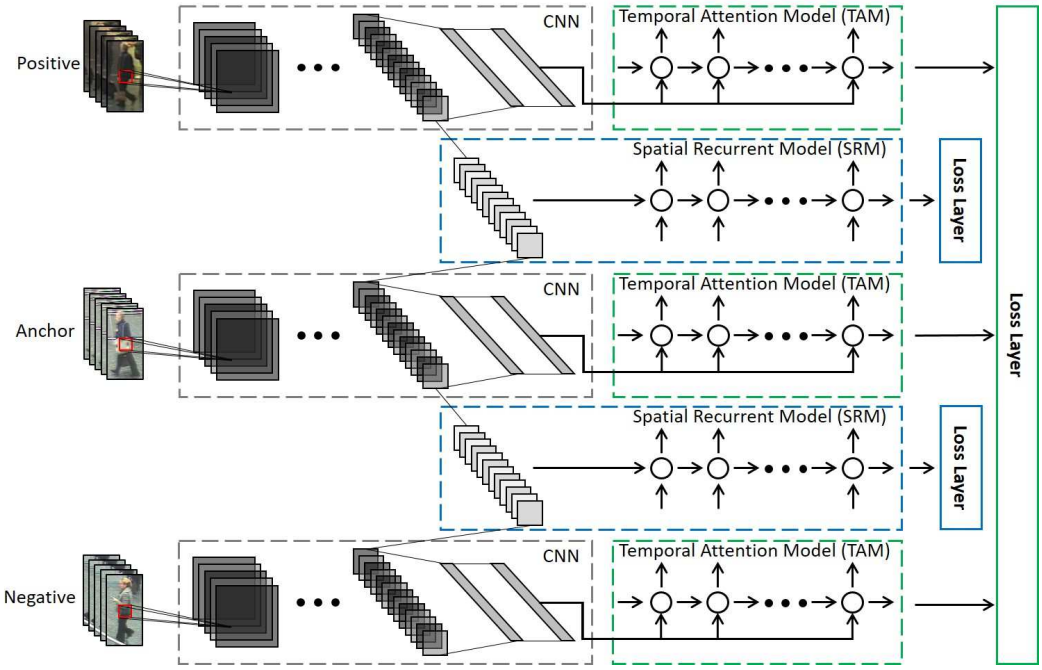

Figure 2: The proposed network architecture. The green boxes are designed for the process of feature learning and the blue ones are for metric learning. The weights of all convolution layers are shared for both processes.

matching task. They mix these two strategies together to learn features and metrics simultaneously and obtain better results than any individual one does.

The proposed method in this paper avoids evaluating frames by hand-crafted features as [29]. It learns the weight of each frame by the temporal RNN. Based on these learned weights, the input of each time step of the temporal RNN is the weighted average of the image sequence. This can be regarded as an implicitly temporal alignment as [22]. As shown in Figure 2, our model is built upon [23, 28]. It accepts a triplet of image sequences as the inputs. After extracting features by a CNN, we apply a temporal RNN to improve feature learning. Meanwhile, spatial RNNs are exploited to learn a good metric. Therefore, the proposed method jointly performs feature learning and metric learning, and integrates both temporal and spatial information at the same time.

2.2. Spatial RNNs and Temporal RNNs

Except those exploiting RNNs for video-based person re-identification [23, 31], there are some other different RNNs in image-based approaches. Liu *et al*. [21] apply an attention model to learn the weight of each pixel in an image. Haque *et al*. [2] use an attention model for the depth data, which learns for localizing a specific region. In this paper, we use a similar attention mechanism as in [35], which attempts to describe a video with proper words. The proposed method selectively focuses on truly relevant frames as shown on the top of Figure 1.

There are several works running an RNN spatially over a feature map. Byeon *et al*. [5] propose an RNN that sweeps horizontally and vertically in both directions across an image. Visin *et al*. [27] take a similar mechanism to consider the surrounding information for semantic segmentation. Bell *et al*. [3] exploit spatial RNNs to compute contextual features for object detection. In this paper, to measure the similarity between two image sequences, we employ spatial RNNs to integrate the surrounding similarities around each location within the same frame and the similarities through contextual frames.

3. The Proposed Method

In this section, we first present the overall architecture of the proposed method, and then explain each of its important components in more details.

3.1. Overall Architecture

Supposing each image sequence is represented as $x = \{x_t | x_t \in \mathbb{R}^D\}_{t=1}^T$. T is the length of the image sequence and D is the dimension of images. As shown in Figure 2, the proposed method accepts a triplet of image sequences as the inputs. In each stream, we first employ a CNN to extract

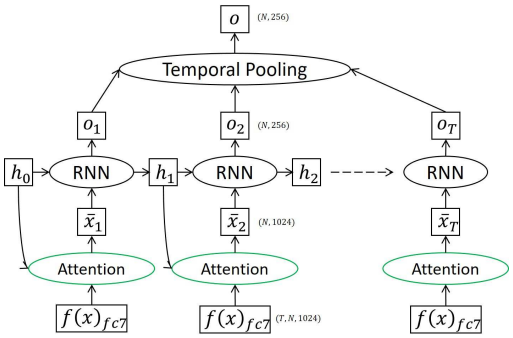

Figure 3: The structure of the temporal attention model. The input is T feature maps of the fc7 layer of the image sequence x. N is the batch size. After an attention sub-net, a weighted average of the T features, $\bar{x}_t$, is obtained. Then $\bar{x}_t$ is fed into an RNN, which outputs the feature o_t at each time step. The final representation of x is the temporal average pooling of $\{o_t\}_{t=1}^{T}$.

the feature of each image x_i. We select CaffeNet[1] for CNN, which is similar to AlexNet [17] except that normalization layers and pooling layers exchange the position. It has five convolution layers (conv1∼conv5) and two fully connected layers (fc6∼fc7). We denote the CNN as $f(x)$ and the feature map of the fc7 layer as $f(x)_{fc7}$.

Then a temporal attention model comes into play to explore temporal features, which comprises a sub-net to learn the relevance of each frame and an RNN to learn feature representations. The RNN can be expressed by

$$g(f(x)_{fc7}) : \mathbb{R}^{T\times D_1} \mapsto \mathbb{R}^{D_2}, \tag{1}$$

where D_1 and D_2 are the dimensions of the fc7 layer and the output of the RNN, respectively. The output of the temporal RNN is denoted by $\mathcal{F}(x)$. For feature learning, the triplet loss [24] is adopted to pull similar pairs and push away dissimilar pairs.

Meanwhile, given a pair of image sequences x^i and x^j, we develop a new stream to separately learn the metric by calculating the element-wise difference between $f(x^i)_{pool5}$ and $f(x^j)_{pool5}$. Then the feature maps will be fed into the spatial recurrent model, which contains six RNNs. Each RNN will sweep the feature map along a specific direction. The output is further processed to generate the final probability that the image sequence pair is of the same person or different people, which is denoted as $M(x^i, x^j)$. In this part, the person re-identification problem is regarded as binary classification task.

During testing, the final similarity between x^i and x^j can

[1]http://caffe.berkeleyvision.org/model_zoo.html

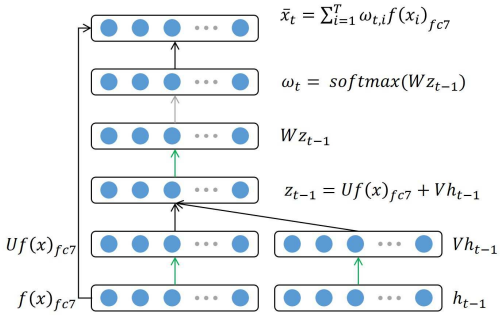

Figure 4: The sub-net to learn the relevance of each image in an image sequence, which is represented by $\omega_t = \{\omega_{t,i}\}_{i=1}^{T}$. The green lines indicate that they are fully connected. The black ones stand for the element-wise sum or inner product. The gray line refers to the softmax operation.

be calculated by

$$S(x^i, x^j) = \frac{1}{1 + F\big(\mathcal{F}(x^i), \mathcal{F}(x^j)\big)} + \lambda M(x^i, x^j), \tag{2}$$

where $F(\cdot,\cdot)$ is a distance measure, which is the normalized Euclidean distance in this paper. λ is the trade off between feature learning and metric learning, which is empirically set to 1 in the experiments.

3.2. Temporal Attention Model (TAM) for Feature Learning

To selectively focus on the most relevant images, an attention mechanism is applied to explore the temporal structure of the given image sequence. The whole process of TAM is shown in Figure 3. It consists of two parts, i.e., the attention unit and the RNN unit. At each time step t, the attention unit accepts $\{f(x_t)_{fc7}\}_{t=1}^{T}$ as the input and generates a weighted average of these features, i.e.,

$$\bar{x}_t = \sum_{i=1}^{T} \omega_{t,i} f(x_i)_{fc7}, \tag{3}$$

where $\{\omega_{t,i}\}$ is learned by a sub-net as shown in Figure 4. h_{t-1} is the hidden state of the RNN at time step $t-1$. $Uf(x)_{fc7}$, Vh_{t-1} and Wz_{t-1} are obtained by the fully connected layers. The softmax operation is used to guarantee that $\sum_i \omega_{t,i} = 1$.

Then $\bar{x}_t$ is fed into an RNN, where the Long Short-Term Memory (LSTM) network [26] is adopted. The LSTM network is able to summarize useful information within a long-range sequence. The final representation of the image sequence is the temporal average pooling of its output at each time [23].

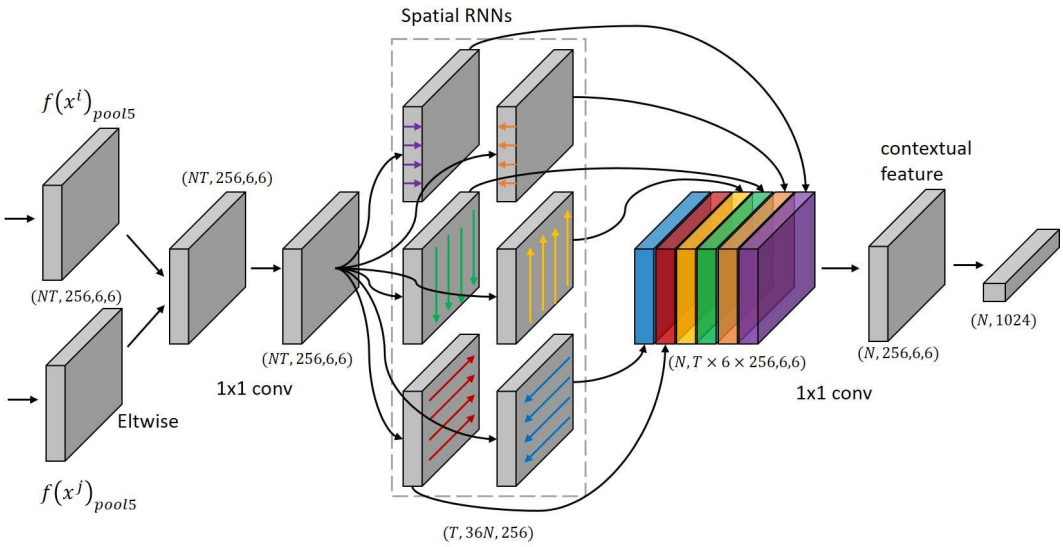

Figure 5: The complete process of the spatial recurrent model for metric learning. There are six spatial RNNs, rendered by different colors. Reshape operations are ignored for better illustration. More details can be found in the context.

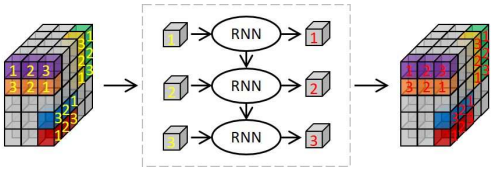

Figure 6: Demonstration of how the spatial RNN works. There are six directions, indicated by different colors. Each volume represents a location in the feature map.

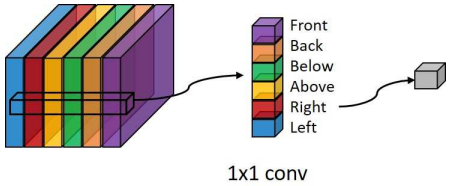

Figure 7: By convolving the stacked outputs of six spatial RNNs with a kernel size of 1×1, we obtain the integrated difference at each location.

3.3. Spatial Recurrent Model (SRM) for Metric Learning

In this paper, the SRM is designed to deal with videos and for metric learning, which contains six spatial RNNs. As pictured in Figure 5, given a pair of inputs, $f(x^i)_{pool5}$ and $f(x^j)_{pool5}$ are mixed together by element-wise minus. The resulting feature map can be regarded as the initial dissimilarity map, followed by a convolution layer with a kernel size of 1×1. Six copies of the feature map are fed into six spatial RNNs, respectively. Each spatial RNN sweeps the feature map along a specific direction as shown in the figure, i.e., forward and backward, from left to right and the opposite, from bottom to top and the opposite. Afterwards, the output of each spatial RNN is stacked together. Another convolution layer with a kernel size of 1×1 follows to summarize the contextual features. A fully connected layers are placed at the end to capture high-order spatial relationships within the contextual features.

Figure 6 illustrates how each spatial RNN works. Each direction is rendered by a specific color. The *left* is the input feature map for spatial RNNs. The numbers in yellow stand for the order along each direction. The *middle* shows the process of an RNN. It accepts the inputs with the given order and generates the outputs with the same order. The *right* expresses that these outputs are placed as the same order as the inputs in the feature map. The LSTM network is selected for RNNs here.

The outputs of six spatial RNNs are then stacked together before a convolution layer with an 1×1 kernel, producing the so-called contextual feature. We further explain the meaning of this convolution. As demonstrated in Figure 7, each location in the stacked feature map stands for the integrated difference along a specific direction. Thus each location in the feature map of the convolution layer is a combination of its six surrounding information. With the help of the proposed SRM, the learned metric will be less sensitive

Datasets	iLIDS-VID	PRID2011	MARS
#identities	300	200	1,261
#track-lets	600	400	21K
#boxes	44K	40K	1M
#distractors	0	0	3K
#cameras	2	2	6
#resolution	64×128	64×128	128×256
#detection	hand	hand	algorithm
#evaluation	CMC	CMC	CMC & mAP

Table 1: The basic information of three datasets to be used in our experiments.

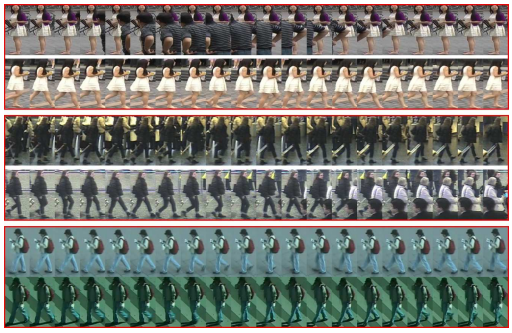

Figure 8: Samples of three datasets used in experiments. The first row shows images from MARS. The following two rows are sampled from iLIDS-VID and PRID2011, respectively.

to illumination changes and occlusions.

4. Experiments

We evaluate our proposed method on three public video datasets. The first part is to verify the effectiveness of the proposed method and its components. And then we compare our method with the state-of-the-art methods. The experimental results demonstrate that the proposed method can enhance the performance of both feature learning and metric learning and outperforms previous methods.

4.1. Datasets

The basic information of three datasets is listed in Table 1 and some samples are displayed in Figure 8.

The iLIDS-VID dataset [29] comprises 600 image sequences of 300 subjects. Each image sequence has a variable length ranging from 23 to 192 frames, with an averaged number of 73. This dataset is challenging due to clothing similarities among people and random occlusions.

The PRID2011 dataset [11] consists of 385 identities in camera A and 749 in camera B. 200 identities appear in both cameras, constituting of 400 image sequences. The length of each image sequence varies from 5 to 675. Following [39], sequences with more than 21 frames are selected, leading to 178 identities.

The Motion Analysis and Re-identification Set (MARS) [39] is a newly released dataset for video-based person re-identification. There are 1,261 pedestrians who are captured by at least 2 cameras. The bounding boxes are generated by a DPM detector [10] and a GMMCP tracker [8]. Among 20,715 track-lets, 3,248 distractor track-lets are produced due to false detection or tracking.

4.2. Implementation Details

We select caffe [14] to implement experiments. CaffeNet is adopted for CNN and LSTM for RNN. The length of an image sequence is experimentally set to 6. The image sequence is randomly selected in a track-let. The dimensions of the fc6 layer and fc7 layer are set to 1,024.

We follow the same protocol [29] for the iLIDS-VID and PRID2011, namely both datasets are evaluated over ten train/test partitions. Each partition splits the dataset into two equivalent parts, one for training and the other for testing. The same experimental setup [39] is adopted for the MARS, i.e., 625 subjects for training and the others for testing. And there are totally 1,980 predefined track-lets in the gallery set. The pedestrians in the training set and testing set are non-overlapping for all three datasets. Images are firstly resized to 227×227 to adjust CaffeNet. We also implement mirror for data augmentation. To accelerate converge, hard negative mining [24] is employed.

Testing a person re-identification system is a ranking problem. Given a query in camera A, we need to calculate the similarity between the query and each candidate in the gallery set in another camera B. The expectation is that the candidates of the same pedestrian in camera B will rank at the top. To evaluate the performance, the Cumulative Matching Characteristic (CMC) [4] curve and Mean Average Precision (mAP) [40] are performed. The CMC curve represents the expectation of the true matching being found within the first n ranks. mAP takes recall into consider when multiple ground truths exist. In our case, CMC and mAP are equvalent for the iLIDS-VID and PRID2011 because they only contain one ground truth in the gallery set while in the MARS multiple ground truths exist. Therefore, both mAP and CMC will be reported for the MARS and CMC is evaluated for the iLIDS-VID and PRID2011.

4.3. Effectiveness of Each Component

Table 2 summarizes the quantitative performance of the baseline methods on the MARS dataset. "CNN" refers to use a CNN to extract features of each frame and measure the similarity by the Euclidean distance. The representation of an image sequence is obtained by using the average

Dataset	MARS			
Rank@R	R = 1	R = 5	R = 20	mAP
CNN	58.5	76.3	85.9	40.3
CNN+RNN	60.3	79.2	87.0	42.0
CNN+TAM	62.7	80.6	90.5	43.4
CNN+DIFF	63.0	81.1	91.5	44.8
CNN+SRM	64.2	84.4	94.3	46.2
ALL	70.6	90.0	97.6	50.7

Table 2: Performance of baseline methods on the MARS dataset.

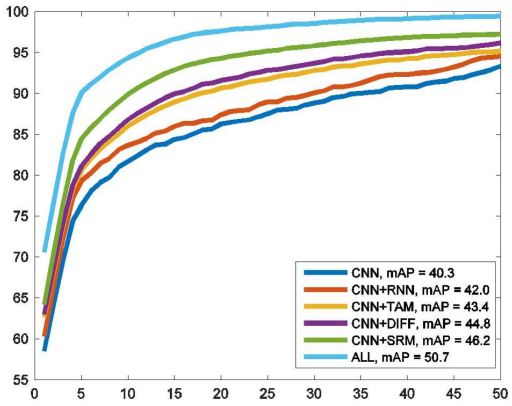

Figure 9: CMC curves of baseline methods on the MARS dataset.

temporal pooling. "CNN+RNN" means that instead of using temporal pooling, an RNN is applied to further process the features and generate the representation of the image sequence. "CNN+TAM" is on top of "CNN+RNN" by exploiting the temporal attention model. "CNN+DIFF" is to directly uses a fully connected layer instead of the spatial RNNs within the SRM after the CNN. "CNN+SRM" is beyond "CNN+DIFF" by using the spatial recurrent model. "ALL" is the proposed full architecture as pictured in Figure 2. Figure 9 shows the CMC curves of these baseline methods on the MARS dataset. It is easy to draw the following conclusions from the above experimental results.

1. By comparing "CNN", "CNN+RNN" and "CNN+TAM", we can conclude that the recurrent attention model works, i.e., it can help to pick out relevant frames.

2. "CNN+SRM" performs better than "CNN+DIFF", which tells that the spatial recurrent model helps to learn better metrics.

3. "ALL" performs the best, which indicates that joint feature learning and metric learning is better than performing them separately.

Figure 10 provides four examples of retrieving. The query in the first row is heavily occluded during the image sequence. The image sequence for the second query suffers apparent illumination changes compared with the matched candidates. The third and fourth query image sequences contain multiple evident pedestrians. Our method succeeds in finding correct matched candidates at the top rank in the first three examples, which shows the robustness of the proposed method with respect to occlusions and illumination changes. The last query fails to retrieve the same pedestrian in another camera. The reason may be that the query contains two equivalent identities in the whole image sequence. Our model can not distinguish whether it is the boy or the girl we want to recognize. In fact, our method has found the girl in the first, sixth and seventeenth candidates.

4.4. Comparison with the State-of-the-art Methods

Table 3 summarizes the comparison of our method with the state-of-the-art methods. The results on the iLIDS-VID and PRID2011 are obtained by finetuning a pretrained model on the MARS. Wang *et al*. [29] and Wang *et al*. [30] propose the so-called flow energy to measure the importance of each frame and accordingly select video fragments. Liu *et al*. [22] consider the temporal alignment and build a spatial-temporal representation for each video. Karanam *et al*. [15] propose to learn a dictionary that tries to encode features discriminatingly and solve the problem of viewpoint variations. You *et al*. [37] try to decrease the intra-class difference of nearby positive samples and push away the nearest negative samples. Mclaughlin *et al*. [23] and Wu *et al*. [31] take a similar deep neural network architecture, i.e., a CNN followed by an RNN. Zheng *et al*. [39] employ ID-discriminative Embedding to directly train a classification model. We achieve the best performance on both MARS and PRID2011, and the comparable result on the iLIDS-VID. The reason for the latter may be that Mclaughlin *et al*. [23] use both color images and the corresponding optical flows while we only use color images. In the future, we will try to combine multiple features as the inputs.

5. Conclusion

In this paper, we have proposed an end-to-end deep neural network architecture, which integrates a temporal attention model to selectively focus on the discriminative frames and a spatial recurrent model to exploit the contextual information when measuring the similarity. We carefully designed experiments to demonstrate the effectiveness of each component of the proposed method. In comparison with the state-of-the-art methods, our method performs the best, which shows that the proposed temporal attention model is

Figure 10: Retrieval results of the proposed method in the testing set of MARS. Image sequences in the first column represent the query. The second column contains candidates in the gallery, where a single image stands for an image sequence for visual-pleasing. Candidates with green boxes indicate that they belong to the same pedestrian as the query. While the red boxes refer to wrong matched image sequences. The images with blue boxes imply that they are distractors, which negatively affect the accuracy.

Datasets	iLIDS-VID			PRID2011			MARS			
Rank@R	R = 1	R = 5	R = 20	R = 1	R = 5	R = 20	R = 1	R = 5	R = 20	mAP
Wang *et al*. [29]	34.5	56.7	77.5	37.6	63.9	89.4	-	-	-	-
Liu *et al*. [22]	44.3	71.7	91.7	64.1	87.3	92.0	-	-	-	-
Karanam *et al*. [15]	25.9	48.2	68.9	40.6	69.7	85.6	-	-	-	-
Wang *et al*. [30]	41.3	63.5	83.1	48.3	74.9	94.4	-	-	-	-
You *et al*. [37]	56.3	**87.6**	**98.3**	56.7	80.0	93.6	-	-	-	-
Mclaughlin *et al*. [23]	**58**	84	96	70	90	97	-	-	-	-
Wu *et al*. [31]	46.1	76.8	95.6	69.0	88.4	96.4	-	-	-	-
Zheng *et al*. [39]	53.0	81.4	95.1	77.3	93.5	**99.3**	68.3	82.6	89.4	49.3
Ours	55.2	86.5	97.0	**79.4**	**94.4**	**99.3**	**70.6**	**90.0**	**97.6**	**50.7**

Table 3: Comparison with the state-of-the-art methods. The highest values are shown in boldface. The literatures in the first block are using traditional methods, while the second block contains deep neural network based methods.

useful for feature learning and the spatial recurrent model is beneficial for metric learning.

In recent years, a lot of efforts have been made to improve the performance in person re-identification. However, it is still far from being available for practical applications. Current issues include serious occlusions, heavy illumination changes, non-rigid deformation of human bodies, clothing with similar colors or textures for different persons, and different clothings for the same person. Moreover, it is time to highlight that the largest restriction for the person re-identification research is the lack of very large scale datasets where many practical issues should exist, especially when the deep neural networks become more and more popular. Consequently, our future work lies in collecting as many data as possible, covering scenes as widely as possible.

Acknowledgements

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61572504, 61420106015), Strategic Priority Research Program of the CAS (XDB02070100), and Beijing Natural Science Foundation (4162058). This work is also supported by grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.

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] L. F.-F. Albert Haque, Alexandre Alahi. Recurrent attention models for person identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[3] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *arXiv preprint arXiv:1512.04143*, 2015.

[4] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. The relation between the roc curve and the cmc. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, pages 15–20. IEEE, 2005.

[5] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3547–3555, 2015.

[6] S.-Z. Chen, C.-C. Guo, and J.-H. Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016.

[7] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[8] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4099, 2015.

[9] 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.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[11] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[12] J. Hu, J. Lu, and Y.-P. Tan. Deep transfer metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 325–333, 2015.

[13] Y. Huang, W. Wang, and L. Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems*, pages 235–243, 2015.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[15] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4516–4524, 2015.

[16] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[17] 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.

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[19] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3223, 2013.

[20] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.

[21] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[22] 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.

[23] 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.

[24] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[25] E. Ustinova, Y. Ganin, and V. Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. *arXiv preprint arXiv:1512.05300*, 2015.

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[27] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393*, 2015.

[28] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[29] 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.

[30] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[31] 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.

[32] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016.

[33] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[34] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016.

[35] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4507–4515, 2015.

[36] D. Yi, Z. Lei, S. Liao, S. Z. Li, et al. Deep metric learning for person re-identification. In *International Conference on Pattern Recognition*, volume 2014, pages 34–39, 2014.

[37] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[38] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Transactions on Image Processing*, 24(12):4766–4779, 2015.

[39] 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.

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