TEMPORAL REGULARIZED SPATIAL ATTENTION FOR VIDEO-BASED PERSON RE-IDENTIFICATION

Xueying Wang[1], Xu Zhao[1]*

[1]Department of Automation, Shanghai Jiao Tong University

ABSTRACT

Video-based person re-identification aims at matching video sequences of a person across different camera views. How to explore the abundant appearance and motion information in a video sequence is crucial to tackle this problem. To this end, we first introduce a parameter-free spatial attention module to emphasize the importance of discriminative regions. Then we apply a temporal regularization term on spatial attention to refine corrupted region caused by occlusion and blur. This term allows the attention response at one position in a frame to be related to other frames of the same position. Extensive experiments are conducted on iLIDS-VID and PRID-2011 datasets. The experimental results demonstrate that our approach surpasses the existing state-of-the-art video-based person re-identification methods on iLIDS-VID and PRID-2011.

Index Terms— person re-id, spatial attention module, temporal regularization, focal loss

1. INTRODUCTION

The problem of person re-identification is to recognize the same pedestrians over non-overlapping camera views. It is a critical research due to its extensive applications, such as public security and forensic investigation [1, 2]. Researches on this problem has received increasing attention in recent years. However, this task still remains as a challenging problem, since there exist intricate variations of body poses, lighting conditions, backgrounds and viewing points.

Person re-identification algorithms can be divided into two categories: still image-based approaches and video-based ones. Most existing researches [3, 4, 5] are made to solve the problems on still images, while video-based person re-identification has received less attention. Compared with image-based methods, video-based methods are more suitable in real-world applications. Video-based re-id provides rich appearance information from different frames so that it is more robust to noise. Besides, motion context information, such as gait, could also be used in identifying pedestrians. In this paper we focus on the video-based person re-identification.

For video-based Re-id, it is critical to fully utilize spatial and temporal features of video sequences. Liu et al. [6] align the dynamic appearance of video sequences globally by body part segmentation and gait information. You et al. [7] design a top-push distance learning model to enforce the optimization on top rank matching in person re-identification. In recent years, deep networks have been extensively utilized in Re-id. Neural networks have been employed to learn more robust features and accurate similarity values for image pairs or triplets. McLaughlin et al. [8] utilize Recurrent Neural Network (RNN) to fuse temporal features extracted from frames. The outputs of RNN from every time step are averaged to generate the sole representation of a video sequence. Xu et al. [9] extend the RNN-CNNs architecture by decomposing pooling into two steps: a spatial pooling layer on CNN features to select key regions, and an attentive temporal pooling layer on RNN outputs to select key frames. Recent studies discard RNN and apply a temporal attention mechanism [10, 11, 12], in which they generate quality scores for frame-level features, and adopt attention weighted average pooling to aggregate temporal features.

However, the methods above ignore an important effect of the temporal information, which has great potential in reducing spatial noise. Once the feature representation of one frame is corrupted by noise, instead of discarding this frame, other frames in the video should provide complementary information to recover this corrupted frame. Also, those temporal attention mechanisms above can only discriminate frames by assigning different attention weights to different frames, therefore, they cannot discriminate either different frames in the video sequence or the different body parts in one frame. Besides, most attention mechanisms consist of convolutional layers which require additional parameters.

In an attempt to solve the above problems, we propose a simple yet effective temporal regularized spatial attention framework, as shown in Fig 1. The main novelties of the proposed framework are summarized as follows:

- We introduce a spatial attention layer to discover salient regions of each input frame and it is entirely parameter-free.

* Corresponding Author. This research has been supported by the National Key Research and Development Program of China (Grant No. 2017YFC0806501) and NSFC Program (61673269, 61273285).

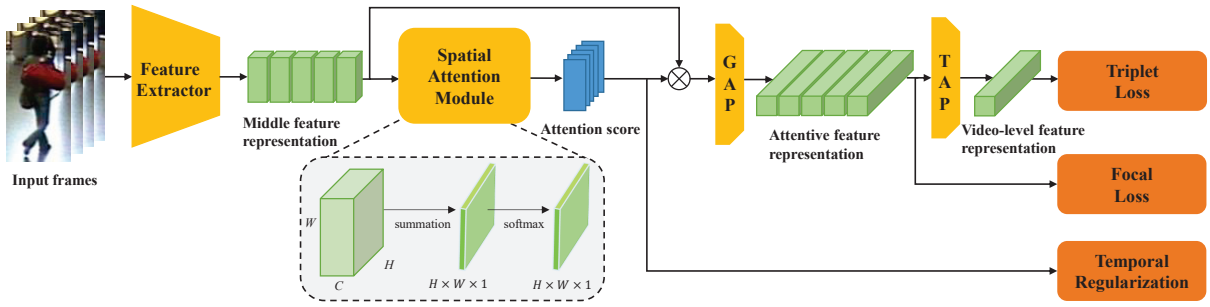

Fig. 1. The overall architecture of the proposed method. GAP: global average pooling. TAP: temporal average pooling.

- We propose a temporal regularization term on spatial attention to learn both discriminative spatial features and discriminative temporal features at the same time. This term is able to eliminate sequence noise and recover corrupted regions across time.
- We demonstrate the effectiveness of our approach on two commonly used person re-id datasets (PRID-2011 [13] and iLIDS-VID [14]). The final results achieve Rank-1 accuracy of 92.1% and 82.7% on PRID-2011 and iLIDS-VID datasets, respectively.

2. APPROACH

2.1. Framework

We choose ResNet-50 CNN [15] as our backbone to extract features of each image. The first layer of ResNet-50 is a convolutional layer called conv1. There are four residual blocks following conv1 named res1, 2, 3, 4 respectively. Some modifications are made on ResNet-50 CNN in this paper. In order to enlarge the final feature map, we set the stride of res5 module to 1. As a result, every input image is represented by a 16×8 feature map of $2048\text{-}dim$. We randomly select N frames in a video sequence, then each video is denoted as $X = \{x^1, x^2, \ldots, x^N\}$. We refer to the modified ResNet as a feature descriptor, $f^n = \phi(x^n)$. Each input frame x^n is represented by f^n after ResNet.

2.2. Spatial Attention Module

Spatial attention mechanism aims at locating discriminative regions. Previous attention modules consist of convolutional layers or fully connected layers which bring high computational cost to training process. In order to simplify the architecture, similar to [16], we introduce a parameter-free spatial attention module followed by the ResNet backbone.

The illustration of the spatial attention module is shown in Fig 1. Given the feature maps of an input video. The size of a feature map f^n is denoted as $H \times W \times C$. First, we sum the feature map along channel dimension thus we get a 2-D feature map R of $H \times W$,

$$R_n(i,j) = \sum_{k=0}^{C} f_k^n(i,j). \tag{1}$$

Instead of using multiple convolutional layers, we generate the corresponding spatial attention map by operating softmax function on R,

$$\alpha_n(i,j) = \frac{e^{R_n(i,j)}}{\sum_{i,j} e^{R_n(i,j)}}. \tag{2}$$

As a result, each position (i,j) is assigned a special attention score $\alpha(i,j)$ which represents importance degree at spatial position (i,j). $\alpha(i,j)$ will be multiplied to all the activations through the depth channel of f^n for the corresponding spatial positions. Therefore, the output v of a parameter-free spatial attention module is written as

$$v^n = f^n(i,j)\alpha_n(i,j). \tag{3}$$

The same operations are applied on all the selected frames of input video.

2.3. Temporal Regularization

For a video sequence, feature representation for each frame should be similar to each other to represent the same person. As mentioned before, for each time step, the attention at each position sum to 1 by the softmax function. However, there is no constraint across time that the attentive receptive field of one frame needs to be similar to the attentive receptive field of another frame. In intelligent surveillance, it is common that some frames in a sequence suffer from occlusion and blur pollution. That makes it possible for the learned spatial attention scores easily detect one specific region and largely ignore other parts of the images so that different frames would locate different regions. In order to alleviate this case, inspired by [17], we impose an additional penalty term over the location attention. To restrict the differences between frames, we need to minimize $1 - \sum_n^N \alpha_n(i,j)$ which

is equivalent to minimizing $(1-\sum_n^N \alpha_n(i,j))^2$. Therefore, the regularization term Reg is defined as

$$Reg = \sum_{i=0}^{H}\sum_{j=0}^{W}(1-\sum_{n}^{N}\alpha_n(i,j))^2. \tag{4}$$

We add this inter-frame regularization term to the original loss function L_{total} defined in Eq.(9) and multiply a coefficient λ,

$$\min(L_{total} + \lambda Reg). \tag{5}$$

2.4. Loss Function

For person re-id, recent studies usually utilize both the ranking loss and classification loss. We employ the triplet loss with hard mining [18] as our video-level ranking loss. This constraint forces the distances of most similar pairs to be smaller than distances of most dissimilar pairs with a margin m. Thus, the ranking loss is defined as

$$L_v = \frac{1}{P}\frac{1}{K}\sum_{i=1}^{P}\sum_{i=1}^{K}[m + \max_{p=1\ldots K} D(v_{i,a}, v_{i,p}) - \min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} D(v_{i,a}, v_{j,n})]_+, \tag{6}$$

where P is the number of identities sampled in a mini-batch and each identity has K tracklets. We define $v_{i,a}$ as anchor samples , $v_{i,p}$ as positive samples and $v_{j,n}$ as negative samples. Here, positive and negative samples indicate samples with the same or different identities from anchor samples, $D(,\cdot,)$ is the L2 distance between two sequence feature embeddings.

Inspired by the impact of hard example mining, we introduce the focal loss [19] proposed in dense object detection on classification task to predict the identity. The idea is to assign hard examples a higher weight than easy examples. Given an image feature representation v_i, the probability of v_i belonging to the c_i-th class is denoted as follows,

$$p_i = \mathrm{Sigmoid_{c_i}}(\mathrm{FC}(v_i)). \tag{7}$$

Then the focal loss can be formulated as follows,

$$L_f = -\frac{1}{PK}\sum_{i=1}^{PK}(1-p_i)^{\gamma}\log(p_i). \tag{8}$$

The total training objective is the combination of focal loss and triplet loss,

$$L_{total} = L_f + L_v. \tag{9}$$

3. EXPERIMENTS

We conduct experiments on two publicly available datasets: iLIDS-VID [14] and PRID-2011 [13]. All experiments are conducted with publicly available code of Pytorch [20] on two NVIDIA GTX1080 GPUs.

3.1. Datasets

iLIDS-VID consists of 600 person sequences for 300 identities. Each identity contains two video sequences captured by two non-overlapping cameras . PRID-2011 includes person sequences from two disjoint camera views, with 385 and 749 persons respectively. However, there are only 200 persons appeared in both cameras. The sequence length of videos in iLIDS-VID range from 23 to 192 frames, with 73 frames as the average duration. In contrast, each sequence in PRID-2011 has 100 frames on average, with sequence length varying from 5 to 675 frames.

3.2. Settings

Following the evaluation protocol in [21]. we randomly split iLIDS-VID and PRID-2011 datasets into two equal-sized sets for training and testing. Cumulative Matching Characteristics (CMC) [22] and mean Average Precision (mAP) are adopted to evaluate our method proposed in this paper.

As discussed in section 2, we randomly select N=15 frames from each input sequence, and adopt the modified ResNet pretrained on the ImageNet [23] dataset as our backbone. The input images are resized to 256×128 pixels. In order to increase data diversity, random horizontal flipping is applied. We employ temporal average pooling as our feature aggregation function so that a variable-length input sequence can be mapped to a video-level representation of fixed dimension. The triplet loss margin is recommended to set at 0.3, the coefficient λ in Eq.(5) is set to 1 and γ in Eq.(8) is set to 2. Adaptive Moment Estimation (Adam) with a weight decay of 0.0005 is deployed as our optimizer. During training process, we set the total number of epochs to be 600, starting with a learning rate of 0.0003 and decay to 3×10^{-5} and 3×10^{-6} rate after 200 and 400 epochs. During inference, ranking loss, classification loss and regularization term are discarded, and Euclidean distances between extracted features are used to measure sequence similarities.

3.3. Ablation Study

We analyses the effectiveness of each module of our proposed methods. We carry out several experiments including w/ or w/o triplet loss, w/ or w/o spatial attention module, w/ or w/o temporal regularization and focal loss or cross-entropy loss. All our experiments on iLIDS-VID and PRID-2011 datasets are of the same settings as discussed above. In Table 1, the performance of each component is listed.

Table 1. Ablation study on each component of our model on iLIDS-VID and PRID-2011 datasets. Rank 1, 5, 10, 20 accuracies (%) and mAP (%) are reported.

Model	iLIDS-VID					PRID-2011				
	R1	R5	R10	R20	mAP	R1	R5	R10	R20	mAP
Baseline	61.2	85.3	91.2	94.5	74.0	84.3	96.6	97.2	98.9	87.4
Baseline + L_v	68.6	88.7	90.4	94.8	77.7	88.0	97.2	97.8	99.5	89.7
Baseline + L_v + SAM	73.5	91.4	92.0	96.0	81.2	91.0	97.5	98.5	99.8	92.3
Baseline + L_v + SAM + Reg	81.3	94.1	96.0	98.7	86.2	91.3	98.9	99.9	100.0	93.2
Baseline + L_f + SAM + Reg	82.7	95.5	98.7	99.3	87.9	92.1	98.8	100.0	100.0	94.1

Baseline refers to modified ResNet trained with softmax cross-entropy loss, in which the video sequence length N is set to 15. Average pooling is adopted to generate the single video-level representation. $\mathbf{L_v}$ represents hard-batch triplet loss, and "$+L_v$" means hard-batch triplet loss combined with softmax loss. **SAM** is the spatial attention module we introduced. It produces a 16×8 attention score map and this map is used to calculate the weighted feature map of each input frame. Compared with $\mathbf{Baseline + L_v}$, **SAM** improves Rank-1 accuracy by 4.9% on iLIDS-VID and 3% PRID-2011 respectively. From the results, we can see that spatial attention module is very effective at detecting discriminative regions which is helpful for boosting re-identification performance. **Reg** corresponds to the proposed temporal regularization term. After adding **Reg** to $\mathbf{Baseline + L_v + SAM}$, the Rank-1 accuracy and mAP improve by 7.8% and 5.0% on iLIDS-VID, as well as 0.3% and 0.9% on PRID-2011 respectively. It is obvious that this temporal regularization term can alleviate frame diversity and recover noise part thus the re-id performance is improved. $\mathbf{L_f}$ means replacing the cross-entropy loss in $\mathbf{Baseline + L_v + SAM + Reg}$ with focal loss. As shown in Table 1, on iLIDS-VID, focal loss exceeds cross-entropy loss by 1.4%/1.7% in Rank-1/mAP, respectively. And in PRID-2011, the improvment reaches 0.8%/0.9% in Rank-1/mAP. It is obvious that all the components of our model contribute to our final results. When all the components are added together, representing as $\mathbf{Baseline + L_f + SAM + Reg}$, we achieve 82.7% and 92.1% Rank-1 accuracy on iLIDS-VID and PRID-2011 datasets respectively. Meanwhile, for each input sequence, only 0.0672 seconds are needed to generate the feature representation.

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

We compare the performance of our proposed framework with some state-of-the-art methods. In Table 2 and Table 3, we show the recognition rates at Rank 1, 5, 10, 20 respectively for both iLIDS-VID and PRID-2011 datasets. On iLIDS-VID dataset, we attain the highest performance for Rank-1 accuracy, that is 82.7%. Compared with method STAN, the Rank-1 improvement achieved by our approach is 4.1%. On PRID-2011 dataset, our approach is only slightly lower than STAN in terms of Rank-1 accuracy. This is because our approach cannot be fully expressed under simple backgrounds and few occlusion. The results verified that our approach achieves superior performance over other video-based state-of-the art methods.

Table 2. Performance comparison of proposed model with the state-of-the-arts on iLIDS-VID dataset in terms of Rank 1, 5, 10, 20 matching rate (%). The best results are marked bold.

Methods	R1	R5	R10	R20
STA [6]	44.3	71.7	83.7	91.7
TDL [7]	56.3	87.6	95.6	98.3
RNN [8]	58	84	91	96
ASTPN [9]	62	86	94	98
QAN [10]	68.0	86.8	95.4	97.4
RQEN [11]	76.1	92.9	97.5	99.3
STAN [12]	80.2	-	-	-
Ours	**82.7**	**95.3**	**98.7**	**99.3**

Table 3. Performance comparison of proposed model with the state-of-the-arts on PRID-2011 dataset. The best results are marked bold.

Methods	R1	R5	R10	R20
STA [6]	64.1	87.3	89.9	92.0
TDL [7]	56.7	80.0	87.6	93.6
RNN [8]	70	90	95	97
ASTPN [9]	77	95	99	99
QAN [10]	90.3	98.2	99.3	100.0
RQEN [11]	92.4	98.8	99.6	100.0
STAN [12]	**93.2**	-	-	-
Ours	92.1	**98.8**	**100.0**	**100.0**

4. CONCLUSION

This paper addresses the video-based person re-identification problem with the proposed Spatial Attention Module and Temporal Regularization term. Also, different from existing works, we introduce a parameter-free attention mechanism and apply hard mining on classification loss. Experimental results demonstrate that our model outperforms other state-of-the-art methods.

References

[1] Shun Zhang, Jinjun Wang, Zelun Wang, Yihong Gong, and Yuehu Liu, "Multi-target tracking by learning local-to-global trajectory models," *Pattern Recognition*, vol. 48, no. 2, pp. 580–590, 2015.

[2] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 3, pp. 334–352, 2004.

[3] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[4] Rui Zhao, Wanli Oyang, and Xiaogang Wang, "Person re-identification by saliency learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 356–370, 2017.

[5] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*. Springer, 2014, pp. 1–16.

[6] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *ICCV*, 2015, pp. 3810–3818.

[7] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng, "Top-push video-based person re-identification," in *CVPR*, 2016, pp. 1345–1353.

[8] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, 2016, pp. 1325–1334.

[9] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *ICCV*. IEEE, 2017, pp. 4743–4752.

[10] Yu Liu, Junjie Yan, and Wanli Ouyang, "Quality aware network for set to set recognition," in *CVPR*. IEEE, 2017, pp. 4694–4703.

[11] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai, "Region-based quality estimation network for large-scale person re-identification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *CVPR*, 2018, pp. 369–378.

[13] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.

[14] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang, "Person re-identification by discriminative selection in video ranking.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2501–2514, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[16] Haoran Wang, Yue Fan, Zexin Wang, Licheng Jiao, and Bernt Schiele, "Parameter-free spatial attention network for person re-identification," *arXiv preprint arXiv:1811.12150*, 2018.

[17] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[19] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[21] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang, "Person re-identification by video ranking," in *ECCV*. Springer, 2014, pp. 688–703.

[22] Douglas Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*. Citeseer, 2007, vol. 3, pp. 1–7.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.