MULTI-SCALE SPATIAL-TEMPORAL NETWORK FOR PERSON RE-IDENTIFICATION

Zhikang Wang, Lihuo He, Xinbo Gao, Yuanfei Huang

School of Electronic Engineering, Xidian University, China

ABSTRACT

Video-based person re-identification (ReID) is an important task, which has received much attention in recent years due to its efficiency in the field of surveillance. Researchers have employed many effective approaches for video-based person ReID, but there are still two problems. Firstly, the same pedestrian in the video sequences differs in size. Secondly, traditional RNNs can only process one-dimension features, which are not suitable for dealing with video sequences. To solve above problems, we propose a new network called Multi-scale Spatial-Temporal Network (MSTN), which combines multi-scale feature extractor and CLSTM together to tackle the discrepant sizes of pedestrians and extract more representative temporal information for the video sequences. We conduct the experiments on the iLIDS-VID, PRID-2011 and MARS datasets, and our approach outperforms state-of-the-art methods by a large margin.

***Index Terms*—** Multi-scale features, CLSTM, temporal information, person re-identification

1. INTRODUCTION

With the development of the monitoring system, person ReID plays a more significant role than ever before in our daily life. More and more people have begun research on it, resulting in its blowout development. The task of person ReID is to re-identify the same person appearing from other cameras. Due to the intensive changes, such as light, pose, viewpoint and occlusions, person ReID is still a very challenging task.

Recent years, researchers propose many image-based person ReID algorithms, which include GAN learning [1, 2], distance metric learning [3, 4], etc. Apart from image-based person ReID, there exists video-based person ReID [5, 6, 7, 8] as well. While comparing the differences between images and video sequences, we have the intuition that video sequences have more temporal information. Moreover, video sequences contain much samples of persons' appearances, resulting in more discriminative learned features. Besides, video-based dataset has a large number of samples, making it much easier to train the machine learning algorithms and deep neural networks.

The main idea of video-based methods is to extract useful representations for video sequences. Then a distance function is learnt to find the matched person. When considering the video based person ReID, we find that the size of the same pedestrian in the video sequences is various, which motivates us to solve the problem with multi-scale features to tackle the discrepant sizes of pedestrians. Additionally, video sequences have rich temporal information. Through the RNNs, we can learn the connections between the images in video sequences.

In this paper, we propose the Multi-scale Spatial-Temporal Network (MSTN), a powerful mechanism for learning the representative features of the video sequences. Specifically, the extractor is constructed to learn multi-scale features, which are used for the first stage training. Then, we extract the temporal information through the CLSTM blocks. Finally, we adopt the classification model instead of siamese model for the network.

2. RELATED WORK

Person ReID is a challenging task, which has been widely researched for many years. Due to the large variations of lighting conditions, viewing angles, body poses and occlusions, many approaches have been applied to solve the problem. In this section, we firstly describe some approaches about multi-scale network. Then we discuss RNNs for video-based person ReID. Finally, we will introduce a new model named CLSTM.

2.1. Multi-scale network

Due to the diversity of pedestrians' sizes in images, many researchers would like to deal with the problem with multi-scale network [9, 10, 11, 12]. In [9], Liu *et al.* proposes a multi-scale triplet convolutional neural network, which captures visual appearance of a person at various scales. They optimize the network parameters by a comparative similarity loss on massive samples triplets. In [12], Wang *et al.* propose an end-to-end feature learning strategy which integrates discriminative information with various granularities. They carefully design the Multiple Granularity Network (MGN), a multi-branch network architecture consisting of one branch for global features and two branches for local features. However, very few researchers try to solve the video-based person ReID problem through the multi-scale network. Pedestrians in video sequences are different scales due to the limitation

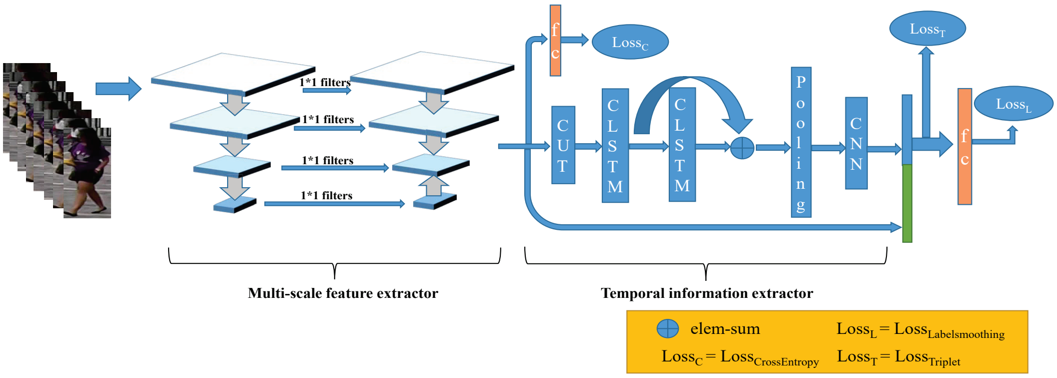
1*1 filters
1*1 filters
1*1 filters
1*1 filters
f c
Loss$_C$
Loss$_T$
Loss$_L$
CUT
CLSTM
CLSTM
Pooling
CNN
f c
Multi-scale feature extractor
Temporal information extractor
elem-sum
Loss$_L$ = Loss$_{Labelsmoothing}$
Loss$_C$ = Loss$_{CrossEntropy}$
Loss$_T$ = Loss$_{Triplet}$

Fig. 1. Our Multi-scale Spatial-Temporal Network for video-based person ReID. We adopt the classification network architecture and three losses for training the network.

of detection algorithms, resulting the multi-scale problem of the network needs solving urgently. To solve the problem of diverse sizes, we proposed our own multi-scale network for extracting the multi-scale features.

2.2. RNNs for video-based person re-identification

Different from image-based person ReID, a video sequence has more temporal information than a single image, which makes it a more nature way to perform person ReID. More and more people use RNN to extract the temporal information [6, 7, 13]. In [7], McLaughlin *et al.* propose a novel Recurrent Neural Network architecture for video-based person ReID, which allows information to flow between time-steps. In [13], Xu *et al.* propose joint spatial and temporal attention pooling network, which combines CNN, RNN and attention mechanism for person ReID and integrates features with temporal information successfully.

2.3. CLSTM

Considering that the input of traditional RNNs must be a one-dimension vector, which is not quite suitable for image-based processes such as person ReID. Therefore, we would like to adopt a new generalized LSTM, denoted by CLSTM, which was widely used in [14, 15, 16]. CLSTM replaces the dot product of LSTM by convolution. This means the input can be two-dimension feature vectors. It is particularly efficient in solving the problem of video sequences. Specifically, CLSTM can be formulated as follows:

$$\begin{cases} i_t = \sigma(x_t * W_{xi} + h_{t-1} * W_{hi} + b_i) \\ f_t = \sigma(x_t * W_{xf} + h_{t-1} * W_{hf} + b_f) \\ c_t = c_{t-1} \odot f_t + i_t \odot \tanh(x_t * W_{xc} + h_{t-1} * W_{hc} + b_c) \\ o_t = \sigma(x_t * W_{xo} + h_{t-1} * W_{ho} + b_o) \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{1}$$

Here, $\sigma()$ and $\tanh()$ are logistic sigmoid and hyperbolic tangent functions; i$_t$, f$_t$, o$_t$ are the input gate, forget gate, and output gate. b$_i$, b$_f$, b$_c$, b$_o$ are bias terms. x$_t$, c$_t$, h$_t$ are the input, the cell activation state, and the hidden state respectively. W$_{**}$ are weight matrices to control the value transitions of the parameters. For instance, W$_{hf}$ controls how the forget gate takes values from the hidden state and W$_{xo}$ controls the transition from input to output. $*$ denotes convolution and $\odot$ denotes element-wise product. As we can see, CLSTM has the same formulation as classic LSTM except the operators.

3. THE PROPOSED ARCHITECTURE

In our work, we build a Multi-scale Spatial-Temporal Network (MSTN) for video-based person ReID. Our MSTN architecture works by passing the video sequences through a classification network to extract the features, which are more representative than others. As shown in Fig.1, each sequence is imported into a multi-scale CNN network to extract features. Then those features are transferred to subsequent network and function Loss$_c$. After that, we use the CLSTM blocks to extract the intra temporal connection of the video sequence. The features before the fc layer and feature vectors of fc layer will be transformed to calculate the Loss$_t$ and Loss$_l$ respectively. Finally, we will get the representative features after training.

The crucial part of MSTN architecture relies on the multi-scale feature extractor and CLSTM blocks. We will introduce details about the network in the following subsections.

3.1. Multi-scale feature extractor

Our multi-scale feature extractor is based on the ResNet [17] backbone. We define the output features of residual blocks conv2∼5 as C2, C3, C4 and C5 respectively. As shown in Fig.1, 1 $*$ 1 convolution filters are applied on C2, C3, C4. C5 to cut the dimension of the features. Most methods just use global features like C4 or C5 for subsequent calculations. However, the inability to distinguish between the foreground and the background is a disadvantage of the global features itself. In person ReID, pedestrians are occluded from time to time. At this time, local features like C2 and C3 are needed. The local features refer to some points that can appear stably and distinguishable. Thus, in the case where the pedestrian is not completely occluded, some local features are still stable to represent the object. At the same time, we also need the global features like C4 and C5 to provide more semantic information of the images. Then, we adopt average pooling for C2, C3 and up sampling for C5 to make sure features extracted from conv2∼5 have the same size. We concatenate the four parts of features as f1, which will be send into a fc layer to calculate Loss$_c$ for adjusting the former part's parameters. We adopt the Softmax Loss which can be formulated as:

$$L_{soft\max} = -\sum_{j=1}^{T} y_i \log \frac{e^{a_i}}{\sum_{K=1}^{T} e^{a_k}} \tag{2}$$

In the function, a$_j$ is the j$_{th}$ element in the feature vector and T is length of feature vector. y$_i$ is the label of the pedestrian.

3.2. Temporal information extractor

Considering the amount of model's calculation, it is necessary to cut the dimension of the multi-scale features. So, we apply a block named 'cut', which is a 1*1 convolutional layer, to set the dimension of each frame to 512. Due to the intrinsic connection of adjacent pixel points, we should avoid directly straightening the feature and inputting it into the LSTM network for training. Compared with dot product, convolution is a better way to process images. Therefore, we adopt CLSTM blocks to extract temporal information in the network. With the network depth increasing, the accuracy gets saturated and then degrades rapidly. To solve the degradation problem, the feature extracted by the first CLSTM block will be added to the second CLSTM block's output. We use the max pooling to extract the max value of each channel and apply the one-dimension convolution to get the temporal features. We then concatenate the temporal features with f1 as the final features. Finally, the cross entropy loss with label smoothing regularizer [18] and a triplet loss function [19] are used for further training. We randomly sample K clips for P identities (each clips contains T frames), totally PK clips in a batch. The network will select the hardest positive and the hardest negative samples for each sample in the batch to form the triplets. Then, we will calculate the triplet loss, which can be formulated as:

$$\begin{aligned} L_{triplet} = & \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\ anchors} [m + \overbrace{\max_{p=1...K} D(f_{i,a}, f_p^i)}^{hardest\ positive} \\ & \underbrace{- \min_{j=1..P,n=1..K,j\neq i} D (f_{i,a}, f_{j,n})}_{hardest\ negative}]_+ \end{aligned} \tag{3}$$

Table 1. Performance of each component

Dataset	MARS				
Rank@R	R=1	R=5	R=10	R=20	mAP
CNN	52.1	69.7	76.8	81.7	35.2
CNN+LSTM	70.1	85.0	89.6	92.9	56.5
CNN+CLSTM	74.1	87.8	91.2	93.2	60.5
CNN+CLSTM+Triplet	78.1	91.0	94.0	95.1	70.1

After the fc layer, we train the network by label smoothing cross entropy loss:

$$L_{labelsmoothing} = (1 - epsilon) * y + epsilon/k \tag{4}$$

where epsilon is hyperparmeter we defined, k is the number of classes. We use the final vector of fc layer and the same equation as Equation.2 to calculate y. As the training processing, the network tends to pay more attention to the inner connection of the sequences, paying less attention to the poor detection images. The final loss of our network can be formulated as:

$$Loss = Loss_{soft\max} + Loss_{triplet} + Loss_{labelsmoothing} \tag{5}$$

4. EXPERIMENTAL RESULTS

In this section, we evaluate proposed model for video-based person ReID on three different datasets: iLIDS-VID [22], PRID-2011 [23] and MARS [20]. We also summarize the effects of each component in our model MSTN respectively.

4.1. Datasets

iLIDS-VID and PRID-2011: The iLIDS-VID and PRID-2011 datasets contain 300 and 200 pedestrians respectively. All the sequences are observed by two disjoint camera views in public space. Video sequences in iLIDS-VID have variable length ranging from 23 to 192, with an average number of 73. The length of video sequences in PRID-2011 is from 5 to 675 and with an average of 100.

MARS: The MARS dataset is the biggest video-based person ReID dataset so far. It contains 1261 IDs and around 20000 tracklets. Every person is captured by at least two cameras and has an average of 13.2 tracklets. The dataset also contains quite a few false detections such as buildings, making it in line with reality. Compared with the first two datasets, MARS dataset is much more challenging. Therefore, we focus our main search on MARS datasets.

4.2. Experiment Settings

For iLIDS-VID and PRID-2011 datasets, we randomly divide the dataset into two equal parts at the initial of the experiment, one part is for training and the other part is for testing. We

Table 2. Performance of methods on three dataset

Dataset	iLIDS-VID			PRID-2011			MARS			
Rank@R	R=1	R=5	R=20	R=1	R=5	R=20	R=1	R=5	R=20	mAP
Xu *et al*.[13]	62	86	98	77	95	99	44.0	64.0	81.0	-
Zhang *et al*.[8]	39.5	66.9	86.5	68.5	84.7	96.3	56.5	70.6	79.8	-
Zheng *et al*.[20]]	53.0	81.4	95.1	77.3	93.5	99.3	68.3	82.6	89.4	49.3
Zhou *et al*.[6]	55.2	86.5	97.0	79.4	94.4	99.3	70.6	90.0	97.6	50.7
Zhang *et al*.[21]	60.2	84.7	95.2	85.2	97.1	99.6	71.2	85.7	94.3	-
Liu *et al*.[5]	68.0	86.8	97.4	90.3	98.2	100	-	-	-	-
Ours	**74.0**	**94.0**	**99.3**	85.4	94.4	98.4	**78.1**	**91.0**	**95.3**	**70.1**

make sure that there is no cross between the training and testing part. As for dataset MARS, we strictly follow the protocol released by [17] to divide the dataset. Since the variable length of video sequences, four groups of eight consecutive pictures are selected to represent each pedestrian.

We evaluate the performance of the model through Cumulative Matching Characteristics (CMC) curves and Mean Average Precision (mAP). CMC curves and mAP reflect the search precision and recall ratio respectively.

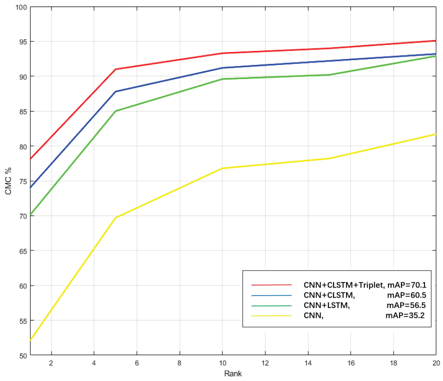

Fig. 2. CMC curves of each component respectively.

4.3. Effectiveness of each component

Table 1 summarizes the effects of each component on the MARS dataset. 'CNN' refers to the use of the multi-scale feature extractor to extract features of each frame, followed by the adaptive average pooling. 'CNN+LSTM' means that the multi-scale feature extractors will be imported into the LSTM block for getting the temporal information of the sequences. 'CNN+CLSTM' is designed on top of 'CNN+RNN' by replacing the LSTM block with the CLSTM. We also add an experiment to train the model with hard triplet loss. Fig.2 shows the CMC curves of different models on MARS dataset. We can draw three conclusions from the above experiment results.

1. After comparing 'CNN', 'CNN+LSTM' and 'CNN+CLST', we can conclude that features with temporal information can work better.

2. 'CNN+CLSTM' does better than 'CNN+LSTM', which implies that CLSTM block is more suitable for processing video sequences.

3. After using the Triplet loss, mAP of the results increases by a large margin, which means Triplet loss can help to increase the recall ratio.

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

In Table 2, we compare our method with other state-of-the-art methods. We train the network on iLIDS-VID, PRID2011 and MARS datasets individually. Xu *et al.* [13] and Zhang *et al.* [8] both adopt a similar architecture, which uses CNN to extract the spatial features and employs RNN to extract the temporal information in video sequences. Zheng *et al.* [20] employ ID-discriminative Embedding to train a classification network directly. Zhou *et al.* [6] builds an end-to-end deep neural network architecture to learn spatial-temporal features and metrics jointly. Zhang *et al.* [21] firstly propose an end-to-end network basing on reinforcement learning. Liu *et al.* [5] propose the network, which can learn the quality of each sample automatically and aggregate the features with the quality score for person ReID. We achieve the best performance on iLIDS-VID and MARS, and the comparable result on PRID-2011. The lack of sufficient training data leads to the bad performance on PRID-2011 dataset. In the future, we will try to simplify the network structure for easy training.

5. CONCLUSION

In this paper, we present a novel Multi-scale Spatial-Temporal Network (MSTN) for video-based person ReID in an end-to-end fashion. In contrast to most existing video-based person ReID methods that ignore the changes in pedestrians' sizes and disadvantages of traditional RNNs, the proposed model is capable of extracting multi-scale features and latent temporal information for video sequences. Experiment results on iLIDS-VID, PRID2011 and MARS datasets show the effectiveness of the proposed method. In the future work, we plan to reduce the model's parameters and propose a lightweight architecture.

6. REFERENCES

[1] Z Zheng, L Zheng, and Y Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017.

[2] L Wei, S Zhang, W Gao, and Q Tian, "Person transfer gan to bridge domain gap for person re-identification," *CVPR*, 2018.

[3] L Ma, X Yang, and D Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3656–3670, 2014.

[4] W Zheng, S Gong, and T Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[5] Y Liu, J Yan, and W Ouyang, "Quality aware network for set to set recognition," in *CVPR*, 2017.

[6] Z Zhou, Y Huang, W Wang, L Wang, and T Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *CVPR*, 2017.

[7] N McLaughlin, J Martinez del Rincon, and P Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, 2016, pp. 1325–1334.

[8] D Zhang, W Wu, H Cheng, R Zhang, Z Dong, and Z Cai, "Image-to-video person re-identification with temporally memorized similarity learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[9] J Liu, Z Zha, Q Tian, D Liu, T Yao, Q Ling, and T Mei, "Multi-scale triplet cnn for person re-identification," in *ACM Multimedia Conference*, 2016, pp. 192–196.

[10] X Qian, Y Fu, and Y Jiang, "Multi-scale deep learning architectures for person re-identification," in *ICCV*, 2017.

[11] Y Chen, X Zhu, and S Gong, "Person re-identification by deep learning multi-scale representations," in *ICCVW*, 2018, pp. 2590–2600.

[12] G Wang, Y Yuan, X Chen, J Li, and Xi Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *CVPR*, 2018.

[13] S Xu, Y Cheng, K Gu, Y Yang, S Chang, and P Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *ICCV*, 2017.

[14] V Patraucean, A Handa, and R Cipolla, "Spatio-temporal video autoencoder with differentiable memory," in *ICLR*, 2016.

[15] X Shi, Z Chen, and H Wang, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015, pp. 802–810.

[16] J Chen, L Yang, and Y Zhang, "Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation," in *NIPS*, 2016, pp. 3036–3044.

[17] K He, X Zhang, S Ren, and J Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[18] C Szegedy, V Vanhoucke, S Ioffe, J Shlens, and Z Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[19] A Hermans, L Beyer, and B Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[20] 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 *ECCV*. Springer, 2016, pp. 868–884.

[21] J Zhang, N Wang, and L Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *CVPR*, 2018.

[22] T Wang, S Gong, X Zhu, and S Wang, "Person re-identification by video ranking," in *ECCV*. Springer, 2014, pp. 688–703.

[23] M Hirzer, C Beleznai, P M Roth, and H Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.