Effective Similarity Measurement for Video-based Person Re-identification

Yiheng Liu, Chao Xie, Wengang Zhou, Houqiang Li
Department of Electronic Engineering and Information Science
University of Science and Technology of China, Hefei, China
lyh156@mail.ustc.edu.cn, {chaoxie,zhwg,lihq}@ustc.edu.cn

***Abstract*—Learning discriminative spatial-temporal feature representation and distance metric is crucial for video-based person re-identification. Most of current approaches directly use the extracted feature vectors to compute similarity, while a single feature vector is not sufficient enough to overcome the noise caused by background clutters as well as larger variations in poses and viewpoints. To this end, we incorporate learning spatial-temporal feature representation and similarity measurement into a unified framework for video-based person re-identification. We propose a similarity measurement layer, which measures the implicit similarity of two video sequences in different regions. This strategy makes the network more robust to noise. Meanwhile, in order to alleviate the imbalance in the number of positive and negative samples, we propose a matching sampling loss to help training the similarity measurement layer. We extensively conduct comparative experiments on three challenging datasets iLIDS-VID, PRID-2011 and MARS. The experimental results demonstrate that the proposed approach can achieve favorable/superior performance compared with the state-of-the-art methods for the video-based person re-identification.**

***Index Terms*—Person Re-identification, Person Retrieval, 3D CNN, Metric Learning, Neural Network**

I. Introduction

Person re-identification aims to identify the same person under different cameras. This task has drawn more and more attention, because of its significant applications for video surveillance such as finding lost children in the supermarket or tracing fugitives.

Impressive progress has been witnessed for image-based person re-identification. Many effective discriminative feature representation learning methods [1], [2] and distance metric learning methods [3] are proposed. In recent years, benefited from the significant appearance representation capability of convolutional neural network (CNN), many deep neural network models [4], [5] have made great achievements. However, background clutters as well as larger variations in poses and viewpoints introduce a lot of noise into the similarity measurement, which weakens the discrimination capability of the feature vectors extracted by networks. So some methods attempt to embed similarity measurement to neural networks in order to form an end-to-end unified framework. In [6], an alignment method is proposed, which calculates the shortest path between two sets of feature maps. The most common method is training a classifier to measure whether two input images belong to the same person. The difference between these methods is the information fed into the classifier. In [7]–[9], the concatenated feature maps of two images are fed into convolutional layers to integrate the information, which is the input of classifier. In [10], the input of convolutional layers is the element-wise difference between feature maps of two images.

Video sequences contain much richer visual information than a single image. What's more, the motion context in video sequences is useful to distinguish different persons. Attracted by the advantages of video sequences, many methods [11]–[15] are designed for video-based person re-identification. Most of these approaches first learn the appearance representation of each frame and then use average pooling or recurrent networks to fuse them, which cannot effectively mine spatial-temporal clues. Features extracted by 2D convolutional neural network are usually sparse and abstractly encode global features. Only a small part of temporal information in video sequences is retained in the high-level feature maps. This limits their capability to capture temporal information.

Different from previous works, we propose a new network architecture, which consists two key modules. The former is the feature extraction module, which simultaneously learns the appearance representation and motion context. This strategy can effectively mine the spatial-temporal information. The latter is the similarity measurement layer (SML), which focuses on the distinctive patches and measures the similarity of two video sequences. Compared with the common similarity measurement strategies, SML achieves higher performance with less parameters. Meanwhile, the proposed matching sampling loss is useful for alleviating the unbalanced data problem during training SML. Through the cooperation between them, the network learns more discriminative spatial-temporal feature representation and more accurate similarity measurement for video-based person re-identification.

II. The proposed approach

A. Feature Extraction Module

Let $\mathbf{I}_i = \{\mathbf{I}_{i,k}\}_{k=1}^{K}$ represent K different video sequences of person i. The input of the network is $\{\mathbf{I}_i\}_{i=1}^{N}$ consisting K video sequences of N persons. Each video sequence contains D frames.

The feature extraction module (3D ConvNet) consists of five 3D convolutional blocks as shown in Fig. 1. Motivated by

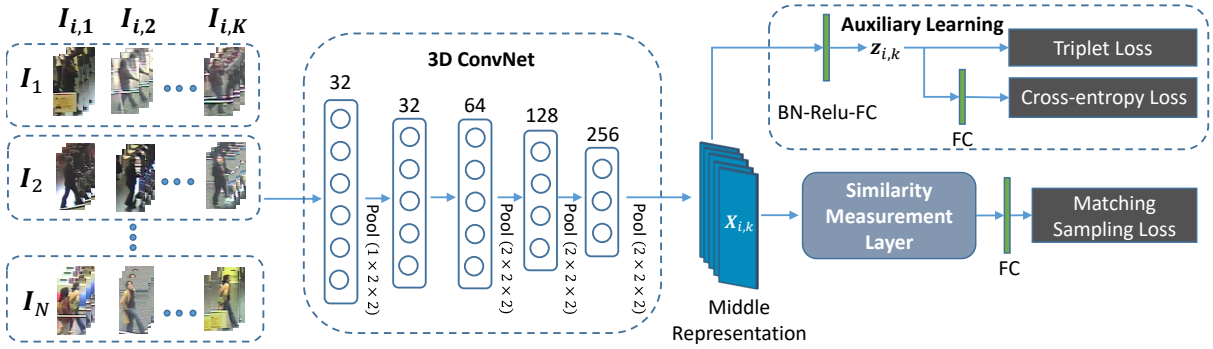

Fig. 1. The overall network architecture of the proposed method. The number above the convolutional block denotes the channel number of the convolutional layer.

[16], each 3D convolutional block is composed of three consecutive operations: 3D batch normalization (BN), followed by a rectified linear unit (ReLU) and a 3D convolutional layer with kernel size of $3 \times 3 \times 3$. Each 3D convolutional block is followed by a 3D max pooling layer, except for the second convolutional block. Given a video sequence input $\mathbf{I}_{i,k}$, we extract its middle feature representation $\mathbf{X}_{i,k} \in \mathcal{R}^{C \times H \times W}$ using the proposed 3D ConvNet, where $H \times W$ is the size of feature maps. C is the number of feature maps. Then, we apply BN, ReLU and a fully connected (FC) layer on $\mathbf{X}_{i,k}$ to get the final feature vector $\mathbf{z}_{i,k} \in \mathcal{R}^{128}$ for auxiliary learning.

B. Similarity Measurement Layer

Given the middle feature maps of two video sequences $\{\mathbf{X}_{i,k}, \mathbf{X}_{j,q}\}$, we concatenate them to form cross-sequence representation $\mathbf{F}_{ik,jq} \in \mathcal{R}^{C \times 2H \times W}$. We use M attention masks to calculate the implicit similarity in channel-wise.

Each attention mask focuses on different regions in a pair of feature maps that are generated by a same convolutional kernel. The output is formulated as

$$s_{ik,jq}^{c,m} = \mathbf{A}_m \odot \mathbf{F}_{ik,jq}^{c}, \tag{1}$$

where $\mathbf{A}_m$ containing $2H \times W$ parameters is the m^{th} attention mask. $\mathbf{F}_{ik,jq}^{c} \in \mathcal{R}^{2H \times W}$ is the feature map in c^{th} channel of $\mathbf{F}_{ik,jq}$. $\odot$ means element-wise multiplication and returns the sum of the absolute value of all elements. Each attention mask can be considered as a similarity measurement function. Then $s_{ik,jq}^{c,m}$ represents the similarity measurement of $\mathbf{F}_{ik,jq}^{c}$ under the similarity measurement function $\mathbf{A}_m$.

The final similarity is formulated as

$$\mathbf{S}_{ik,jq} = \begin{bmatrix} s_{ik,jq}^{1,1} & \cdots & s_{ik,jq}^{1,m} \\ \vdots & \ddots & \vdots \\ s_{ik,jq}^{C,1} & \cdots & s_{ik,jq}^{C,M} \end{bmatrix}. \tag{2}$$

Then $\mathbf{S}_{ik,jq} \in \mathcal{R}^{C,M}$ represents the similarity measurement of C pairs of feature maps under M similarity measurement functions, which is a comprehensive similarity expression of two video sequences. Different attention masks focus on different regions. No matter how the poses change, the activation regions of feature maps always fall into the attention regions of some attention masks, which enhances network's resistance to larger variations in poses and viewpoints.

Given the similarity measurement $\mathbf{S}_{ik,jq}$, we use a FC layer containing 2 softmax units to predict whether these two video sequences belong to the same person. SML and the FC layer form the similarity measurement block (SMB). The parameter number of SML and FC layer is $2 \times H \times W \times M$ and $C \times M \times 2$, respectively. M does not need to be very large, because the spatial resolution of the last feature maps $H \times W$ is small. Several attention masks are sufficient to cover most situations. This means that SMB only introduces a few parameters into the network, which is a great advantage over other similarity measurement strategies. We will provide justification for this in the experiments section.

C. Matching Sampling Loss

In a mini-batch, each video sequence has $K-1$ positive pairs and $(N-1) \times K$ negative pairs. Directly training SMB using cross-entropy loss will induce the network to consider that the input video sequences are always from different persons, because the number of negative pairs is much larger than the number of positive pairs.

Inspired by the batch hard triplet loss [17], which selects the hardest positive and the hardest negative samples within a batch to compute triplet loss, we propose a new matching sampling loss to deal with this problem as a matching task. For each video sequence, we sort the prediction loss values of its positive pairs and negative pairs and pick the top r_p higher classification error from positive pairs and the top r_n higher classification error from negative pairs to optimize the network. The matching sampling loss can be formulated as:

$$\mathcal{L}_m = \sum_{i=1}^{N} \sum_{a=1}^{K} [\sigma_{p=1\ldots K}^{r_p}(\mathbf{I}_{i,a}, \mathbf{I}_{i,p}) + \sigma_{\substack{j=1\ldots N \\ n=1\ldots K \\ j \neq i}}^{r_n}(\mathbf{I}_{i,a}, \mathbf{I}_{j,n})], \tag{3}$$

where $\sigma^{r_p}(\cdot,\cdot)$ and $\sigma^{r_n}(\cdot,\cdot)$ mean computing the classification error and summing the top r_p or r_n prediction loss. Benefiting from the matching sampling loss, we can control the proportion of positive and negative samples. The unbalanced data problem is alleviated and the hard sample pairs are reserved to train the network.

Fig. 2. Attention mask examples of similarity measurement layer. Half of the attention mask is shown, which corresponds to one feature map of a feature map pair. These attention masks focus on the head, limbs, hand-held objects, upper body, lower body and the whole body, respectively.

D. Multi-task Network Architecture

The overall objective function for the proposed network is composed of three parts corresponding to three tasks. As suggested by [6], [14], [15], we use the cross-entropy loss $\mathcal{L}_c$ to optimize classification task. In addition, triplet loss reduces intra-class variations, enlarges inter-class variations and obtains better orders on the given image set. So we adopt batch hard triplet loss [17] to optimize the ranking task, which is formulated as

$$\mathcal{L}_r = \sum_{i=1}^{N}\sum_{a=1}^{K}[m+\max_{p=1\cdots K} D(\mathbf{z}_{i,a},\mathbf{z}_{i,p})-\min_{\substack{j=1\cdots N\\ n=1\cdots K\\ j\neq i}} D(\mathbf{z}_{i,a},\mathbf{z}_{j,n}))]_+, \tag{4}$$

where m is the margin and $D(\cdot,\cdot)$ denotes the Euclidean distance between two feature vectors. As discussed in Sec. II-C, $\mathcal{L}_m$ is adopted to optimize the matching task. The loss function of the whole multi-task network is

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_r + \lambda \cdot \mathcal{L}_m, \tag{5}$$

where $\lambda = 2$ is used to speed up convergence.

In summary, the proposed network architecture consists of three modules: 3D ConvNet, similarity measurement block and auxiliary learning block. Given a video sequences as input, we first use 3D ConvNet to extract the middle representation. Triplet loss and cross-entropy loss are used for auxiliary learning. The final similarities are generated by SMB.

III. Experiments

A. Datasets and Protocols

iLIDS-VID [11] has 300 persons. Each person has two video sequences observed from two non-overlapping cameras views. All sets are used. PRID-2011 [18] contains 400 video sequences for 200 persons observed from two cameras. Following [13], sequence pairs with more than 21 frames are used. MARS [19] includes 1261 persons observed from six cameras. All sequences are automatically generated by pedestrian detector. Sequences with more than 8 frames are used in our experiments.

The MARS dataset has a given partition for training and testing set. The results are reported in terms of Cumulative Matching Characteristic (CMC) table. Following [14], [15], iLIDS-VID and PRID-2011 are randomly split into two sets for training and testing. For testing set, one camera's data is used as probe set, while another camera's data is used as gallery set. The average CMC table over 10 trials with different train/test splits is used to measure the performance of different methods on these two datasets.

TABLE I
Ablation study of the proposed method with different settings on iLIDS-VID. The CMC scores (%) at rank 1, 5, 10, 20 are reported. *3D ConvNet+SMB (WO)* denotes that we use the loss of all positive pairs and negative pairs to constitute final matching loss.

CMC rank	1	5	10	20
2D ConvNet + MeanPooling	62.3	86.7	93.1	96.9
2D ConvNet + MaxPooling	62.7	87.7	94.2	97.4
3D ConvNet (Baseline)	64.1	88.1	94.7	98.8
3D ConvNet + Concat	68.4	90.6	95.7	99.0
3D ConvNet + Diff	63.8	86.8	93.3	97.7
3D ConvNet + SMB (WO)	69.0	90.2	95.93	98.3
3D ConvNet + SMB	71.3	95.1	95.8	98.7

B. Implementation Details

In all experiments, each video frame is resized into 128×64 pixels. Mirroring, cropping and normalization are used to augment datasets. Networks are trained by stochastic gradient descent algorithm with a learning rate of 0.01 and weight decay of 10^{-5}. We set $N = 16, K = 4, M = 64, r_p = 4, r_n = 32, m = 0.4$ in all experiments. Each sub-sequence is consecutive and randomly selected, which contains 8 frames. For testing, in order to take full advantage of the entire video sequence, the average spatial-temporal feature generated by a sliding window with a size of 8 frames and an overlap of 4 frames describes the whole video sequence.

C. Analysis of the Proposed Method

Analysis on fusion strategy. In order to explore the impact of fusion strategy, SMB is not used. The first three networks in Table I are trained by cross-entropy loss and triplet loss. During testing, we directly use cosine distance to compute the similarities of feature vectors $\mathbf{z}_{i,k}$. RNN method based on the proposed CNN architecture does not converge, so the result is not reported. 3D ConvNet achieves the best performance. This indicates that 3D ConvNet can mine more spatial-temporal information than pooling strategy.

Analysis on SMB. We add similarity measurement block (SMB) to the baseline network to form the final network as shown in Fig. 1. Triplet loss and cross-entropy loss are used for auxiliary learning. During testing, the prediction scores represent the similarities of video sequence pairs. The network with SMB achieves 7.2% gains at rank-1 compared with the baseline method. This shows that SML can better measure the similarities of video sequence pairs than directly computing distance based on single feature vector $\mathbf{z}_{i,k}$. Fig. 2 gives some examples of attention masks in SML. We observe that these attention masks focus on different parts of persons. Some pay more attention to local regions (such as head, upper body, limbs, and lower body) while some measure the similarity from the global perspective. Through the cooperation of these attention masks, the proposed network can better handle the

mismatching problem caused by background clutters as well as larger variations in poses and viewpoints.

TABLE II
COMPARISON OF OUR PROPOSED METHOD *3D ConvNet+SMB* WITH EXISTING METHODS AT CMC RANK-1 ACCURACY ON THREE DATASETS.

Datasets	iLIDS-VID	PRID-2011	MARS
CMC rank	**rank-1**	**rank-1**	**rank-1**
TDL [12]	56.3	56.7	-
ASTPN [15]	62.0	77.0	44.0
CNN+RNN [14]	58.0	70.0	40.0
CNN+XQDA [19]	53.0	77.3	**65.0**
DCMA [20]	60.0	80.0	63.0
DGIPR [21]	66.0	74.0	-
3D ConvNet + SMB	**71.3**	**81.5**	61.3

When we remove the matching sampling loss from the proposed method and use the loss of all sample pairs to optimize SMB, the performance drops, which indicates that matching sampling loss is useful for unbalanced matching task.

We also compare SMB with other similarity measurement strategies, *i.e.*, concatenating strategy and difference strategy. The concatenating strategy uses a convolutional layer with 256 filters and a FC layer to predict identities based on the concatenated feature maps of two video sequences, while the difference strategy is based on the difference feature maps of two video sequences. SMB exceeds concatenating strategy 2.9% and exceeds difference strategy 7.5% at rank-1 accuracy. Furthermore, SMB only needs 3.6×10^4 parameters, which is less than concatenating strategies (1.2×10^6) and difference strategy (6×10^5). So SMB achieves higher performance with less parameters. This is because SMB utilizes the prior information that the similarity measurement only needs to be computed for the feature maps generated by the same convolutional kernel, which makes SMB more effective than concatenating strategy. Meanwhile, the different attention masks let SMB be robust to poses and viewpoints changes, while the difference strategy is easily be effected by noise.

D. Comparison with the State-of-the-Art Methods

In order to verify the capability of our approach, we compare the proposed network with the state-of-the-art methods on iLIDS-VID, PRID-2011 and MARS. Results in Table II show that our method performs the best method on both iLIDS-VID and PRID-2011. Compared with the competitive method DGIPR, the improvements achieved by our method are 5.3% and 7.5% at rank-1 accuracy in iLIDS-VID and PRID-2011, respectively. This demonstrates the proposed method can better take advantage of spatial-temporal information and get more accuracy similarity measurement. All sequences in MARS are automatically generated by pedestrian detector. Temporal information has been lost badly, which is not suitable for our method to obtain motion context. Nonetheless, our method still achieves favorable results on MARS.

IV. CONCLUSION

In this paper, we propose a new feature extraction module and a similarity measurement method for video-based person re-identification. The feature extraction module uses 3D ConvNet to mine the spatial-temporal clues, which is more effective than previous methods. What's more, the use of similarity measurement layer trained with the proposed matching sampling loss makes the network more robust to background clutters as well as larger variations in poses and viewpoints. The remarkable results on iLIDS-VID and PRID demonstrate that our method can integrate spatial-temporal information and measure similarity better. Even on MARS dataset that temporal information is corrupted, our approach still achieves favorable results.

REFERENCES

[1] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008, pp. 262–275.

[2] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014, pp. 1–16.

[3] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[4] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016, pp. 791–808.

[5] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016, pp. 1249–1258.

[6] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[7] Y. Zhang, X. Li, L. Zhao, and Z. Zhang, "Semantics-aware deep correspondence structure learning for robust person re-identification." in *IJCAI*, 2016, pp. 3545–3551.

[8] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification." in *AAAI*, vol. 1, no. 2, 2017, p. 3.

[9] C. Mao, Y. Li, Y. Zhang, Z. Zhang, and X. Li, "Multi-channel pyramid person matching network for person re-identification," *arXiv preprint arXiv:1803.02558*, 2018.

[10] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," *arXiv preprint arXiv:1709.05165*, 2017.

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

[12] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *CVPR*, 2016, pp. 1345–1353.

[13] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *ECCV*, 2016, pp. 701–716.

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

[15] 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," *arXiv preprint arXiv:1708.02286*, 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*, 2016, pp. 630–645.

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

[18] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *SCIA*, 2011, pp. 91–102.

[19] 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*, 2016, pp. 868–884.

[20] Z. Song, B. Ni, Y. Yan, Z. Ren, Y. Xu, and X. Yang, "Deep cross-modality alignment for multi-shot person re-identification," in *ACMMM*, 2017, pp. 645–653.

[21] L. Chen, H. Yang, S. Wu, and Z. Gao, "Data generation for improving person re-identification," in *ACMMM*, 2017, pp. 609–617.