MULTI-LEVEL SUPERVISED NETWORK FOR PERSON RE-IDENTIFICATION

Junpeng Zhang, Fei Jiang

Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
jpzhang1995@sjtu.edu.cn, jiangf@sjtu.edu.cn

ABSTRACT

Most existing methods for person re-identification (re-id) only utilize last layer features of deep convolution neural networks (CNNs), which may lose the useful lower level information and decreases the performance. To alleviate this problem, we propose a novel multi-level feature learning framework, Multi-Level Supervised Network (MLSN), for re-id. Specifically, MLSN is consist of a backbone deep CNN and several part-wise sub-networks. The backbone CNN extracts multi-level semantic feature maps at each middle layer and the part-wise sub-networks transform the feature maps to part-based pedestrian descriptors through a designed supervised learning. By fusing these descriptors, MLSN can utilize both low level information and high level information for re-id. Moreover, an effective optimizing strategy is presented to further improve the performance. Experimental results on three large scale datasets Market-1501, DukeMTMC-reid and CUHK03 show that our proposed approach outperforms several state-of-the-art methods.

Index Terms— Person re-identification, Deep learning, Multi-level feature extraction, Multi-level supervision

1. INTRODUCTION

Person re-identification (re-id) is the progress of matching the same person across a set of images taken from different cameras, which is widely used in video surveillance and security. However, due to similar looks among pedestrians, the re-id systems have to recognize the mirror differences, which is still a challenging task.

Most existing methods [1, 2, 3, 4, 5, 6] use deep convolution neural networks to deal with re-id tasks. They either focus on learning discriminative descriptors at the highest layer feature maps [1, 2, 4, 5] or designing a robust metric to directly learn the distance of two high level descriptors [3, 6]. However, the highest level features are not suitable for distinguishing mirror differences among pedestrians. This is because the high-level features usually capture global information of the whole image and ignore the mirror details during the down-sampling operations within deep CNNs. Thus, a method combines both higher level global features and lower level detail features is necessary for re-id.

Although there already exists several works on multi-level feature learning for person re-id, how to effectively extract multi-level information is still an open problem. Zhao et al. [7] and Chang et al. [8] studied on fusing features from different layers, but no explicit supervision was applied in the middle layers. Guo et al. [9] proposed a multi-level similarity network, which is a verification based method with high computation complexity in testing phase. Wang et al. [10] presented a resource aware multiple resolutions network. Their method is learning middle descriptors at four different resolutions, but other layers are ignored.

In this paper, we propose a novel Multi-Level Supervised Network (MLSN) for person re-id, which is consist of a backbone network and multiple sub-networks, as shown in Fig. 1. The backbone network is a deep CNN (e.g. ResNet50 [11]) which produces different semantic feature maps at multiple middle layers. To explore the information hidden among the middle layer feature maps, we design a part-wised subnetwork to extract part-based descriptors. Descriptors at different levels are concatenated together to form a hyper-descriptor, which contains both lower level information and higher level information. Furthermore, an effective optimizing strategy is presented to further improve the performance. Experimental results on several public datasets demonstrate the effectiveness of the proposed method.

Our contribution can be concluded into three parts:

1. We propose a novel Multi-Level Supervised Network (MLFN) which can effectively extract multi-level information for re-id.

2. We propose a novel Part Embedding Sub-network (PEN) which helps MLSN combine part-wise features and level-wise features with computational efficiency.

3. We propose an effective Semi-Detached Training Schedule (SDTS) which further improves the performance of MLSN.

2. OUR APPROACH

In this section, we first describe the overall architecture of the proposed Multi-Level Supervised Network (MLSN). Then,

The work was supported by NSFC (No. 61671290), the Key Program for International S&T Cooperation Project of China (No. 2016YFE0129500), Shanghai Committee of Science and Technology (No. 17511101903), and China Postdoctoral Science Foundation (No. 2018M642019).

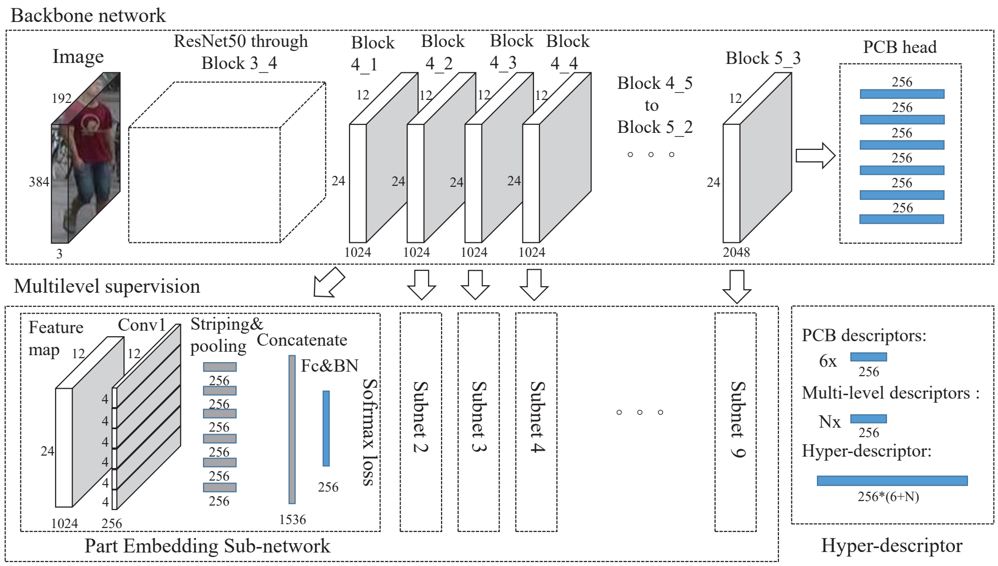

Fig. 1. Illustration of the Multi-Level Supervised Network. MLSN is consist of a backbone network (top part) and multiple re-id sub-networks (bottom part). Particularly, we use ResNet50 with PCB[1] as backbone and totally 9 part-wise sub-network is applied at residual block $B_{4,1}$ to $B_{5,3}$. All the sub-network follow the same structure (bottom left). When testing phase, all descriptors from PCB and some descriptors from sub-networks are concatenated to a hyper-descriptor (bottom right).

we introduce the proposed part-wise subnetwork, called Part Embedding Sub-network (PEN), which is an important component of MLSN. Finally, we focus on how to optimize the MLSN with our proposed Semi-Detached Training Schedule (SDTS).

2.1. Multi-Level Supervised Network (MLSN)

To utilize both low level information and high level information for person re-id, we propose the Multi-Level Supervised Network (MLSN). As shown in Fig. 1, MLSN is consist of a backbone network for feature maps extraction and several sub-networks for multi-level supervised learning. The backbone network can be any deep convolution network, such like Google Inception [12] and ResNet [11]. This paper mainly employs ResNet50 since it is widely used in re-id tasks and it has a natural multi-level architecture. Now we briefly review the ResNet model for simplifying later expression. The ResNet50 model contains 5 convolution stages, denoted as $\{$C1, C2, C3, C4, C5$\}$. Each of the stages is formed by a set of convolution structures named residual block, e.g. stages $\{$ C2, C3, C4, C5$\}$ are formed by $\{$3, 4, 6, 3$\}$ numbers of residual blocks respectively. We use $B_{i,j}$ to denote the j-th residual block at stage i. Given an pedestrian image, each residual block $B_{i,j}$ produces a feature map $F_{i,j}$, while the low-level feature maps tend to capture the detail information and high-level feature maps tend to present the global information. The MLSN learns multiple re-id descriptor based on these multi-level feature maps. Particularly, in this paper, we select $F_{4,1} - F_{4,6}$ and $F_{5,1} - F_{5,3}$, totally 9 feature maps in MLSN. Each feature map is sent into a sub-network to learn a descriptor represents information at this level. The sub-network can be a simple softamx classifier or a metric learning network. In this paper, we apply a custom Part Embedding Sub-network (Sec. 2.2) with softmax classifier. In training phase, all the 9 sub-networks do NOT share parameters and follow a training schedule introduced in Sec. 2.3. Meanwhile, at the top layer we follows the structure of PCB [1] to learn 6 part descriptors. In testing phase, descriptors extracted from sub-networks and PCB heads are concatenated together as the hyper-descriptor of a person. Specially, we select $F_{4,3} - F_{5,3}$ to form the hyper-descriptor according to Sec. 3.4. By measuring the cosine distance among image hyper-descriptors, we can retrieval the most similar pedestrians for re-id.

2.2. Part Embedding Sub-network (PEN)

To learn a more discriminative descriptor at each level, we propose a Part Embedding Sub-network as the sub-network in MLSN. Many researches show that learning part features can significantly improve the accuracy of a re-id system [1, 4]. However in MLSN, directly learning multiple part features at each level is not computation acceptable. Suppose we apply the same method as PCB [1] which divides the features map into six strips and do this partition at ResNet stage $\{$C4, C5$\}$. There would be nine feature maps and each feature map is divided into six parts, totally 54 re-id sub-networks are required for training this multi-level network. Such a big model is not acceptable in computation.

Table 1. Dataset information in our experiments

Dataset	#ID	#Train	#Test	#Image
DukeMCMT-ReID [14]	1402	702	702	36411
Market-1501 [15]	1501	751	750	32668
CUHK03 [16]	1467	767	700	14097

To address this problem, we propose a part-wise sub-network called Part Embedding Sub-network (PEN), which can embed partial features with computation efficient. The structure of PEN is illustrated in bottom left of Fig. 1. Feature map $F_{i,j}$ extracted from ResNet is sent to PEN. An 1x1 convolution layer transform the high dimensional (1024-dim for $F_{4,x}$ and 2048-dim for $F_{5,x}$) feature map into a 256-dim feature map to reduce computation overhead. Then, the reduced feature map is divided into six horizontal strips and each strip is average pooled to a 256-dim vector, which represents the partial feature. After that, all the six 256-dim vectors are concatenated together as a 1536-dim part-based vector. Next, we further reduce the dimension of the 1536-dim part-based vector to a 256-dim descriptor by a bottleneck structure [13] (an additional 256-dim fully connection layer with batch normalization). This 256-dim descriptor is the final output of PEN, which embeds the partial information of a pedestrian image. In training phase, we supervised learn the descriptor by a softmax classifier. In testing phase, the 256-dim descriptor of multiple sub-networks are concatenated together as the final hyper-descriptor of a person. Thus the hyper-descriptor contains both multi-level information and part information.

2.3. Semi-Detached Training Schedule (SDTS)

To better optimize the MLSN, we propose a Semi-Detached Training Schedule. Unlike traditional top-layer supervised training framework, MLSN also does the supervised learning at each middle feature maps, which makes it more difficult to train. Empirically, we found that directly supervise each middle layer decrease the performance of MLSN. One possible reason is gradients from lower layers dominate the training process. Thus, the higher layers can not learn a good descriptor. To overcome this problem, we propose a novel Semi-Detached Training Schedule (SDTS). That means, when doing back propagation in training phase, not all re-id sub-net derive gradients to the backbone ResNet, they are “detached” from the backbone. For the detached sub-nets, back propagation and parameter updating are only done inside themselves. In this paper, we connect the sub-net at residual block $B_{4,6}$ $B_{5,3}$ to backbone and detach all others. Experiment result in Sec. 3.4 show that this strategy benefits the final re-id performance.

3. EXPERIMENT

3.1. Datasets

We conduct experiments on three large-scale datasets: Market-1501 [15], DukeMTMC-reid [14] and CUHK03 [16] to evaluate our method. The detail information of these three datasets is summarized in Table 1. For the evaluation metrics, we adapt standard Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP). For evaluation protocol of CUHK03, we use detected bounding boxes and new protocol proposed in [20].

3.2. Implementation Details

We implement our MLSN model in the PyTorch [21] framework. For the backbone network we use ResNet-50 [11] pretrained on ImageNet. Following [1], We remove the last down sample operation to get higher resolution feature maps. For MLSN structure, we place 9 re-id sub-networks at residual block $B_{4,1}$ to $B_{5,3}$. In each sub-network, we set the parameter 6 for part striping. The output feature of each sub-network is set to 256-dim. At the top layer, we follows PCB [1] structure to learn 6 part descriptors. Thus totally 15 (9 at different levels and 6 at the top level) 256-dim features are output from our MLSN. In training phase, all person images are resized to 384×192. For data augmentation, random horizontal flip, normalization and random erasing [22] is applied. For model optimization, we use stochastic gradient descent. The initial learning rate of backbone network is set to 0.01 and for re-id sub-network and PCB head we set the learning rate 0.1. We train our model with batch size 64, epoch 60, weight decay 5e-4, momentum to 0.9 and learning rate drop 0.1x at epoch 40.

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

Table 2 shows the results of state-of-the-art methods and our methods on three datasets. We divide the results into two groups. The first group is methods based on multi-level features (top). The second group is other high performance methods (middle). Besides, to fairly compare the performance, we denote the methods apply random erasing [22] by (RE). We first compare our model with the first group MT-net [7], DaRe [10] and MLFN [8]. On all the three datasets, our model outperforms all the previous multi-level model in both Rank-1 accuracy and mAP by large margin. Then, comparing with other state of the art methods, our model gets the best performance on DukeMTMC-reid and CUHK03. On DukeMTMC-reid, we surpass the best compared result by +0.5% Rank-1 accuracy and +0.1% mAP. On CUHK03, We surpass the best compared result by +2.1% Rank-1 accuracy and +0.9% mAP. On Market-1501, our method is also comparable with other SOTA methods.

Table 2. Comparision with state-of-the-art methods

Methods	DukeMTMC-reid		Market-1501		CUHK03	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
MT-net [7]	-	-	82.0	63.0	-	-
DaRe [10]	75.2	57.4	86.4	69.3	55.1	51.3
DaRe(RE) [10]	79.1	63.0	88.5	74.2	61.6	58.1
MLFN [8]	81.0	62.8	90.0	74.3	52.8	47.8
HA-CNN [5]	80.5	63.8	91.2	75.7	41.7	38.6
DNN_CRF [17]	84.9	69.5	93.5	81.6	-	-
PCB+RPP [1]	83.3	69.2	**93.8**	81.6	63.7	57.5
GP-reid (RE) [18]	85.2	72.8	92.2	81.2	-	-
Mancs (RE) [19]	84.9	71.8	93.1	**82.3**	65.5	60.5
baseline(PCB) [1]	81.9	65.3	92.4	77.3	61.3	54.2
MLSN	84.5	69.3	91.7	78.8	63.2	58.1
MLSN+PEN	85.2	70.5	92.6	79.1	64.4	58.8
MLSN+SDTS	84.2	70.9	93.1	80.7	64.6	59.3
MLSN+PEN+SDTS	85.4	71.7	93.8	80.5	63.7	58.1
MLSN+PEN+SDTS (RE)	**85.7**	**72.9**	93.2	81.4	**67.6**	**61.4**

Table 3. Rank-1 accuracy for different number of levels. (MLSN+PEN+SDTS)

dataset	feature	top	+1	+2	+3	+4	+5	+6	+7	+8	+9
DukeMTMC-reid [14]	fuse	84.4	84.7	85.0	85.1	85.2	85.1	85.3	**85.4**	85.1	84.8
	single	-	80.8	82.5	**82.8**	81.4	80.3	78.7	75.6	73.2	68.7

3.4. Ablation Studies

In this section, we first analyze the effectiveness of MLSN framework, Part Embedding Sub-network and Semi-Detached Training Schedule. Then we will discuss how to make a balance between low level features and high level features.

Effectiveness of MLSN, PEN and SDTS: To validate our methods, we compare five results: PCB: the baseline model. MLSN: Using a $1 \times 1 \times 2048$ convolution layer with a 256-dim fc-BN layer to replace PEN. MLSN + PEN: MLSN with PEN re-id sub-network . MLSN + SDTS: MLSN with Semi-Detached Training Schedule. MLSN + SDTS + PEN: The full model. The bottom of Table 2 shows the result of each setting. Firstly, we can see that MLSN significant improves the performance than the PCB baseline by +2.6% Rank-1, +4.0% mAP in DukeMTMC-reid; +1.5% mAP in Market-1501 and +1.9% Rank-1, +3.9% mAP in CUHK03, which demonstrate that multi-level supervision is essential for re-id task. Secondly, PEN and SDTS also benefits the result. The PEN makes +0.7%, +0.9%, +1.2% in Rank-1 and +1.2%, +0.3%, +0.7% in mAP on three datasets. SDTS get -0.3%, +1.4%, +1.4% in Rank-1 and +1.6%, +1.9% +1.2% in mAP. Finally, the full model makes the best average performance, especially on DukeMTMC-reid and Market-1501.

How to fuse the hyper-descriptor: In this section we study how many levels is suitable for the final hyper-descriptor. we train a MLSN with 9 re-id sub-networks. Then, in testing phase, we compare different hyper-descriptors concatenated from different number of levels (ResNet Block $\{B_{5,3}\}$; $\{B_{5,3}, B_{5,2}\}$;...; $\{B_{5,3}, B_{5,2}, ..., B_{4,1}\}$ respectively). Table 3 shows the Rank-1 accuracy of these descriptors. In the table, "+i" means the i-th level features from top layer,"fuse" means the feature is concatenated up to i-th level and "single" means the feature at i-th level. We can see that, the Rank-1 accuracy increases while we combines more different levels of features and the best result appears at around 7th levels (+1.0% Rank-1 than 1-level baseline). We also notice that even the low level features can not reach high accuracy singly, it still significant improves the performance when combining with high level features, which we think is because these features contains different semantic information.

4. CONCLUSION

We introduce a novel Multi-Level Supervised Network (MLSN) for person re-id. MLSN can easily capture different level semantic features with flexibility. Combining with our Part Embedding Sub-network and Semi-Detached Training Schedule, MLSN gets state-of-the-art result on three large scale datasets. Our study suggests multi-level features are as important as part features, which has got less attention in previous research. We hope our proposed MLSN can be a useful baseline for future researches on multi-level features learning for person re-id.

5. REFERENCES

[1] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *The European Conference on Computer Vision*, 2018.

[2] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang, "Deeply-learned part-aligned representations for person re-identification.," in *The IEEE International Conference on Computer Vision*, 2017, pp. 3239–3248.

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

[4] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian, "Glad: global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference*, 2017, pp. 420–428.

[5] Wei Li, Xiatian Zhu, and Shaogang Gong, "Harmonious attention network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, vol. 1, p. 2.

[6] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang, "End-to-end deep kronecker-product matching for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6886–6895.

[7] Cairong Zhao, Kang Chen, Zhihua Wei, Yipeng Chen, Duoqian Miao, and Wei Wang, "Multilevel triplet deep learning model for person re-identification," *Pattern Recognition Letters*, 2018.

[8] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang, "Multi-level factorisation net for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, vol. 1, p. 2.

[9] Yiluan Guo and Ngai-Man Cheung, "Efficient and deep person re-identification using multi-level similarity," *arXiv preprint arXiv:1803.11353*, 2018.

[10] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger, "Resource aware person re-identification across multiple resolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8042–8051.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[13] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang, "Svdnet for pedestrian retrieval," in *The IEEE International Conference on Computer Vision*, 2017, pp. 3820–3828.

[14] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *The IEEE International Conference on Computer Vision*, 2017, pp. 3774–3782.

[15] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *The IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[17] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang, "Group consistent similarity learning via deep crf for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8649–8658.

[18] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus, "Re-id done right: towards good practices for person re-identification," *arXiv preprint arXiv:1801.05339*, 2018.

[19] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *The European Conference on Computer Vision*, 2018, pp. 365–381.

[20] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li, "Re-ranking person re-identification with k-reciprocal encoding," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3652–3661.

[21] 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.

[22] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.