Person Re-identification with Joint-loss

Junqi Liu, Na Jiang, Zhong Zhou, Yue Xu
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
zz@buaa.edu.cn

***Abstract*—Person re-identification is a technique that search the given target in the video surveillance network. This technique has been widely applied to security and surveillance system, and also become a research hotspot in computer vision. Person re-identification has been challenging due to the large number of cameras in the network and variation in camera angles, illumination, occlusion and poses. In this paper, we proposed a person re-id approach that can resist occlusions and variations based on a human pose guided convolution neural network framework with joint loss functions. We extract local features from body parts localized by landmarks, merge it with global features to learn the similarity metric. Identification loss and pose-constrained triplet loss function are jointly employed to train the model. Our approach outperforms most state-of-the-art methods on three large-scale datasets, with an accuracy of 83.31%, 86.1% and 72.6% on Cuhk03, Market1501 and Duke MTMC-reID respectively.**

***Keywords*-Person re-identification; Deep learning, Joint loss, Pose estimation**

Figure 1. The challenging of person re-identification.

I. Introduction

Person re-identification (re-id) is a recognition task that associates a given individual across disjoint cameras at different time instances. This technique enables cross-camera tracking and has been widely applied to security surveillance and detection system. Automated person re-id system would be a good substitute for the manual system in terms of both accuracy and efficiency especially in large-scale dataset. However, this task is still challenging due to the variations in camera angles, illuminations, occlusion, poses and etc. As shown in Fig. 1, the images in column (a) to (d) represent the same person but they undergo different variation conditions. (e) is an example of the two images having similar but actually captured from different individuals. All above issue need to be tackle when dealing with a person re-identification task.

Most current approaches of person re-id focus on either features representation or metric learning. Traditional re-id features such as SDALF [4] and Mid-Level Filter [26] have been designed to enhance the robustness against appearance. Unfortunately, due to the occlusion and variation in illumination and pose, mere hand-crafted features have been found incapable for discriminating the people. The application of deep learning has achieved a breakthrough in computer vision. In recent studies, convolutional neural networks such DeepReid [10], DGD [27] have introduced deep learning to feature extraction, leading to a significant increase in accuracy. Nevertheless, these approaches ignore local information thus sensitive to occlusion. To eliminate the influence of occlusion, we propose a novel method that combines local and global information to discover the complementary correlations. Inspired by CPM [6], which estimates human landmark automatically, we localize body parts based on key landmarks, and thereby we are able to capture the local information and combine it with global information. When the feature representation is completed, we need to measure the similarity between every two features. The basic idea of metric learning is that measure the similarity between input image pairs, and find a distance space in which features vectors from same person being closer than those from different ones. This idea has been employed in many verification models like Gated[7], TPC[14], Quadruplet[17] and etc. In our proposed deep learning re-id framework, our experiment has shown that features extracted from identification model has better performs than verification model in intra-class. Meanwhile, features extracted from verification model have has better performs than identification model in terms of inter-class. Therefore, we implement the combination of identification loss and verification loss to train the framework, expecting to obtain the features that are both representative and discriminative. The remainder of this paper is organized as follows: the related work and a general overview of the proposed

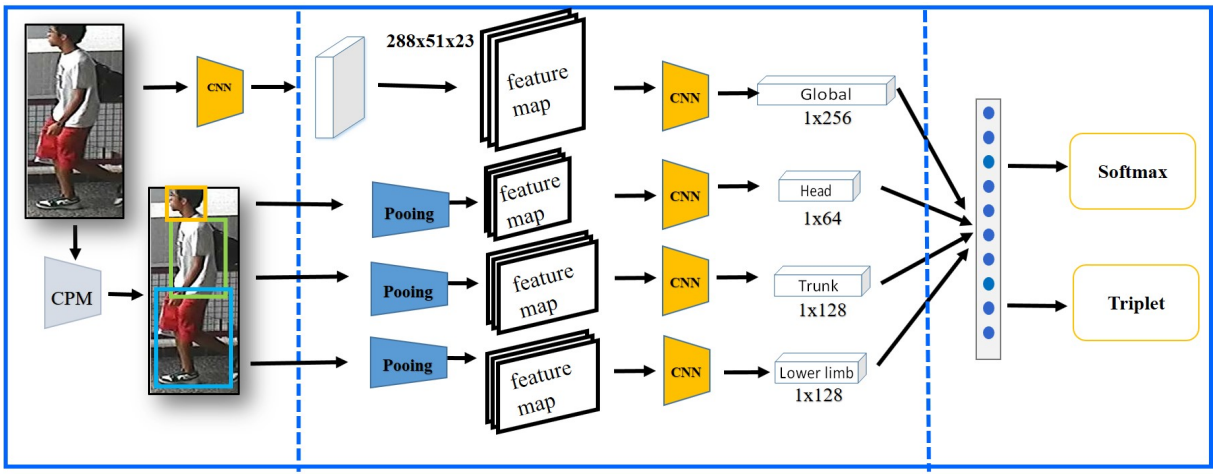

Figure 2. Outline of our framework. It consists of a Convolutional Pose Machine (CPM) with fixed parameters that detects landmarks, a backbone network, 3 branches, 1 concatenate layers for feature fusion and 2 loss layers with joint loss functions for metric learning.

method are presented in Section II. Section III describes the proposed pose guided neural network with joint loss function method for person re-identification. Experiment evaluation including comparison with current method and effectiveness evaluation are discussed in Section IV, and the conclusions are drawn in Section V.

II. Related work

In the past two decades, the re-id problem has become a hotspot in computer vision. According to their research ideas, works fall into two categories. The first category mainly focuses on suitable description of human appearance characteristics. The second category focuses on optimizing distance metric to determine whether a pairwise image represent the same individual. With the deep learning algorithm being widely applied in the voice and visual. Using framework based on the deep neural network for re-id have gradually becomes the mainstream method.

A. Feature Representation

Many efforts have been made to learn features that are both discriminative and invariant to influences. These features are supposed to include both low-level visual information such color histogram, contexture and local features[4] and mid-level semantic description[26].

Most re-id method using handcraft features to represent pedestrian appearance before deep learning has been adopted in person re-id. Among these method, mid-level filter is designed to discover the local salient feature, which achieved Top 1-75.3

In recent years, deep learning has been adopted to solve person re-id, DeepReID[10] employ a CNN model for person re-id. Wang proposed a novel approach Domain Guided Dropout (DGD) to discard useless neurons for each domain and greatly improved the performance on multiple person re-identification datasets. However, mismatch caused by pose change and occlusions has not been resolved. Because this deep framework only focus on global features of the raw images while ignores the spatial and local information. In retrospect of traditional re-id method, most re-id features encode spatial information with different decomposition schemes such as horizontal stripes Gray[1], Kviatkovsky[2],Zheng[3], body parts based on symmetry-driven accumulation of local features. Recent research benefit from the availability of large-scale person re-id dataset[10,11], brings deep learning to feature extractions. Cheng [14] et al proposed a Multi-Channel Parts-Based CNN which jointly learnt global full-body and local features from the original input and horizontal stripes. However, horizontal stripes are sensitive to influences like camera angles and poses, thus introduce error to the model. In 2016, Automatic estimating human landmark algorithm (CPM) is proposed, making it possible to get precise partial location in pedestrian images. Therefore, we utilize the localization of landmarks to promise the semantic alignment, enabling the fusion of local and global features in futures steps.

B. Distance Metric Learning

The second category mainly focuses on metric learning which learning the Mahalanobis distance function that maps features space to distance space that preserves the distance relationship of the given similar/dissimilar input pairs[15]:

$$d(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) \quad (1)$$

where x_i and x_j are the features vectors of sample i and j respectively, M is a semi-defined matrix derived from training samples.

Weinberger proposed LMNN[9] which employed triplets to ensure kNN of each training sample have the same label.

Table I
SKETCH MAP OF OUR BACKBONE NETWORK STRUCTURE.

Module	Output	Channel	1x1	3x3	5x5	1x7	Ave Pool	Max Pool
Data	219x103	3	-	-	-	-	-	-
conv1 3	53x25	64	-	3	-	-	-	1
conv2_x	51x23	192	4	2	1	-	1	-
conv3_x	51x23	288	4	2	1	-	1	-
conv4_x	51x23	288	4	2	1	-	1	-
conv5_x	25x11	768	3	1	-	-	-	1
conv6_x	25x11	768	4	-	-	3	1	-
conv7_x	25x11	768	4	-	-	3	1	-
conv8_x	25x11	768	4	-	-	3	1	-
conv9_x	12x5	1280	2	-	-	1	-	1

Kos[25] designs the equivalence constraints from a statistical inference perspective.

Taking advantage of rapid development in deep learning, researchers tend to combine metric learning with deep learning framework as a verification model. In contrast to the above classification algorithms which regard every individual as an independent, these methods focus on the similarity a pair of images, determining whether they belongs the same individual. Verification model simultaneously receives several inputs such as the image pair [7], triplet or quadruplet [14, 17], these frameworks extract the discriminative feature by network and map them to a distance space in order to measure the similarity among input images. Verification model has advantage to discover discriminate feature due to its definition. However, the frameworks using verification model alone pay more attention to discriminative features but neglect the specific features. Hence the framework using identification model often outperform that using verification model alone. It can be inferred that multi-classification models. It can be inferred that both the specific feature and discriminate feature are significant the person re-identification. In this paper, identification model and verification model are combined in our proposed neural network framework, conceiving to complement each other and improve metric learning.

III. RE-ID WITH JOINT LOSS FUNCTION

In this section, we demonstrate our method from three aspects. First, we localize the body part to estimate poses from human landmarks. Next, we combine global and local features extracted from our proposed network consisting of 1 backbone network and 3 branch networks for final feature representation. Finally, the training process is formulated by jointly minimizing identification loss and pose-constrained verification loss. The body part localization and pose are inferred from human landmarks. Our framework structure has a backbone network and branch network extracting global and local feature respectively. Then the two kinds of feature are merged as the final represent vector, finally we train the model by joint identification loss and pose-constrained verification loss.

A. Partial Localization

CPM is applied to localizing 18 landmarks including neck, nose, shoulders, elbows, wrists, knees, eyes, ears and etc., which provides an estimation of the height of the pedestrian. Thereby, we mark the boundary of body parts with the maximal and minimal coordinates to address local features.

Fig 3 demonstrates how our local bounding box (a) outperforms the fixed ratio horizontal stripes (b) from TPC[14]. Given inaccurate full-body bounding box, horizontal stripes would result in misalignment, while our semantic based decomposition is more relevant and it reduces the redundant background information as well. Even in some extreme cases shown in the third pictures where there are only a few landmarks available due to occlusions, we can still localize body parts with pre-defined regions, which are relatively accurate in most cases.

Figure 3. The result of decomposition by landmarks and horizontal strip.

B. Global and Local Feature Extraction

This section shows the structure of our neural network and the extraction and fusion of global feature and local features.

1) Network Architecture: Inspired by the idea of inception-v3[16], we introduce five different convolutional modules into the backbone network, where each modules

Table II
SKETCH MAP OF OUR BACKBONE NETWORK STRUCTURE.

	global	head	trunk	lower limb
Ave pool	1x1 1280	1x1 640	1x1 1280	1x1 1280
fc	1x1 256	1x1 64	1x1 128	1x1 128

contains several branches stacked by a sequence of multi-scale convolutional and pooling layers. Each row in Table 1 illustrates the schema of the corresponding module, where AveP and MaxP represent the average pooling layer and maximal pooling layer respectively. Taking advantage of this structure, our proposed the network with fewer parameters, is wider and more scale-adaptive. ReLu adds non-linearity to the network. The application of Batch Normalization layers before each ReLu layer accelerates the convergence process, and avoids tweaking the initializations of weights and biases manually. Beside, we randomly dropout 50% neurons of the Fully-connected layer to avoid over-fitting.

Many existing deep learning algorithms use the ImageNet pre-trained CNN models to obtain good score. These model often has square shape numerous parameters. However, pedestrian images are typically small and rectangular, which is not appropriate with the square shape networks. Therefore we proposed a network structure that well fits the pedestrian aspect ratio. The proposed input size is conceived to extracted effective feature and reduce the parameter of network.

2) Fusion of Global and Local Feature: Through the pose estimation, we localize the region of interest (RoI) with bounding box loc^i=(x^i,y^i,w^i,h^i), for body part i where $i \in (1, 2, 3)$ represents head, trunk and lower limb respectively. Each bounding box is defined by its top-left corner (x^i,y^i) and its height and width (w^i,h^i), and thereby mapping the features from the input image by scaling.

Since the size of detected RoIs are flexible, we add a RoI pooling layer to convert the features inside local and the output of conv4_x into a feature map with fixed size. For example, to get an mn feature map from wh input, RoI max pooling works by max-pooling the values in each $w/mh/n$ sliding sub-windows into the output grid cell. The branches share the parameters prior to conv_5 with the backbone network. The pooling layer extracts the local features given the location of RoI. As shown in Table 2, the structure of branches and backbone network are similar, except that branches have fewer outputs from the last pooling layer and fully connected layer, which aids weight tuning. We use local features from connect layer in conjunction with global features during training process. Most existing re-id studies are based on classic framework such as GoogleNet[18] and ResNet50[23], which has good performance in terms of generalization, but typically relies on the pre-trained model loaded on ImageNet[22] to accelerate convergence. The structure of the pre-trained model are fixed, thus less flexible. Benefit from the novel architecture, our model is able to converge fast without pre-trained model.

C. Joint Loss Function

In this joint loss function model, Softmax loss function emphasizes the classification of pedestrian images while pose-constrained triplet loss function determines whether two images contain the same person.

1) Identification Loss Function: In classic identification model, the last fully connected layer is followed by a softmax layer with k outputs, which outputs a probability distribution over the k class. The learning problem in identification network is formulated in terms of minimizing cross-entropy loss that is the identification loss.

$$L_c(v, t, w_i) = \sum_{i=1}^{n} p_i log\hat{p} = -log\hat{p}_t \tag{2}$$

where v is the vector in the final feature layer, representing the features of the image, t is the category, w_i is the parameter in the softmax layer, measuring the difference of target probability distribution p_i and predicted probability distribution $\hat{p}_t$.

2) Triplet Loss Function: Here is the introduction to triplets loss function: I^a,I^p,I^n is defined as a triplet, where I^a could be any reference pedestrian image in the dataset, I^p is another image of the same individual, denoted as positive sample, and I^n is an image of any other individual, denoted as negative sample. f(I_i^a),f(I_i^p),f(I_i^n) can be computed through forward step. We obtain the triplet constraint as follow:

$$D_{id}(I_i^a, I_i^p I_i^n) = d(f(I_i^a) - f(I_i^p)) - d(f(I_i^a) - f(I_i^n)) < \alpha \tag{3}$$

where the function $d(x, y)$ represents Euclidean distance between x and y. The first and second terms on the right-hand side of the equation measure the distance D_{id}. Through solving the inequation, we lean the metric that preserve the distance relationship among image triplets, that is features from the same individual is more similar than those from different individuals.

(a) CUHK03 (b) Market-1501 (c) Duke

Figure 4. The result of decomposition by landmarks and horizontal strip.

IV. EXPERIMENT

This chapter includes the introduction of the datasets we used, experiments on several detests and the evaluation of our approach.

Table III
SKETCH MAP OF OUR BACKBONE NETWORK STRUCTURE.

method	CUHK03		Market-1501		Duke	
	rank1	rank5	rank 1	mAP	rank1	mAP
BoW+KISSME[21]	24.30	45.0	44.42	20.76	25.13	12.17
SDALF[4]	4.90	21.0	20.53	8.20	-	-
S-LSTM[19]	57.3	80.10	61.60	35.30	-	-
Gated[7]	61.80	88.10	65.88	35.68	-	-
GAN[12]	73.1	92.7	78.06	56.23	67.68	47.13
Quadruplet[17]	75.53	95.15	-	-	-	-
SSM[20]	76.63	94.59	82.21	**68.80**	-	-
OIM[8]	77.5	-	82.1	-	68.1	17.04
SVDNet[24]	81.8	-	82.3	62.1	76.7	56.8
ACRN[5]	62.63	89.69	83.61	62.60	72.58	51.96
Ours	**83.31**	**97.50**	**86.10**	67.97	**72.62**	**52.88**

A. Datasets and Evaluation Methods

We evaluate our proposed approach on 3 large-scale benchmark datasets, CUHK03[10], Market-1501[11] and Duke MTMC-reID[12]. Fig. 6 shows some samples in these datasets.

B. Comparison with State of the Art Methods

In this chapter, we compare our result to the existing studies. There are still several disadvantages of this algorithm.

As demonstrated in Table 4, our proposed approach outperform most existing methods, achieving 83.31%, 86.10%, 72.62% rank-1 accuracy on CUHK03, Market1501 and Duke respectively.

Quadruplet [17], ACRN [5] also improve the loss function. In Quadruplet, new negative samples are introduced to triplets to boost metric learning. Basically, it still adopts triplets and measures the distance between positive and negative samples. Unlike Quadruplet, our approach employs an improved triplet loss in conjunction with identification loss function to better discriminate between inner class features. ACRN integrates attribute-complementary information into the triplets to distinguish between input pairs, but the learning process of ACRN is relative complex because the dataset needs to be attribute-labeled before training.

Our approach surpass most state-of-of-the-art researches, except for SSM who shows a 0.83% higher mAP only on Market1501. This is because SSM performs reranking to the result recalled in the first iteration, which precede the correct result and thereby increase mAP.

C. Effectiveness of Joint Feature and Loss

To verify the effectiveness of our approach, we conduct experiment on Market1501, and compare the results from different optimization and parameter setting. Table 5 shows the result.

The baseline experiment only involves backbone network and identification model. The rest are optimization on the baseline setting. Table 5 illustrated that the joint features and joint loss improve the rank-1 accuracy by 3.41% and 5% respectively compared to the baseline. The model adopts triplet and softmax loss weighted exceed other models. Therefore, we assert that the result would improve when higher weight is assigned to the identification during training the model with joint loss. Since the training with joint loss function is fine-tuned based on the baseline which is able to discriminate between inter class features, higher weighted verification model would boost inner class discrimination. The last row of Table 5 shows an integration of all the modification proposed in this paper. The result improves with more optimization scheme added to the model, illustrating the effectiveness of our approach.

Table IV
EFFICIENCY ANALYSIS OF OUR PROPOSED ARCHITECTURE.

	weight ratio of loss	rank-1	mAP
baseline	-	79.72	58.94
Global+Local	-	83.13	67.04
Softmax+Triplet	1:0.5	84.57	67.71
Softmax+Triplet	1:1	84.78	67.87
Softmax+Triplet	0.5:1	84.81	67.23
Whole Architecture	0.5:1	**86.10**	**67.97**

V. CONCLUSION

In this paper, we propose a deep neural network framework for person re-identification in video surveillance. We extract local features from body parts localized by landmarks, merge it with global features to learn the similarity metric. Identification loss and improved triplet loss function are jointly employed to train the model. We evaluate our approach on three benchmark person re-id datasets C, M, D, achieving 83.31%, 86.10% and 72.62% rank-1 accuracy

respectively. This result outperforms most state-of-the-art research. Nonetheless, pose estimate would be unreliable when there are severe occlusions, future work will concentrate on extracting features robust against severe occlusion.

ACKNOWLEDGMENT

This work is supported by the National 863 Program of China under Grant No.2015AA016403 and the Natural Science Foundation of China under Grant No.61572061, 61472020.

REFERENCES

[1] Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features[J]. Computer Vision-ECCV 2008, 2008: 262-275.

[2] Kviatkovsky I, Adam A, Rivlin E. Color invariants for person reidentification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(7): 1622-1634.

[3] Zheng W S, Gong S, Xiang T. Person re-identification by probabilistic relative distance comparison[C]//Computer vision and pattern recognition (CVPR), 2011 IEEE conference on. IEEE, 2011: 649-656.

[4] Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 2360-2367.

[5] Schumann A, Stiefelhagen R. Person Re-Identification by Deep Learning Attribute-Complementary Information[J].

[6] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.

[7] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[C]//European Conference on Computer Vision. Springer International Publishing, 2016: 791-808.

[8] Xiao T, Li S, Wang B, et al. Joint detection and identification feature learning for person search[C]//Proc. CVPR. 2017.

[9] Weinberger K Q, Blitzer J, Saul L K. Distance metric learning for large margin nearest neighbor classification[C]//Advances in neural information processing systems. 2006: 1473-1480.

[10] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 152-159.

[11] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116-1124

[12] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[J]. arXiv preprint arXiv:1701.07717, 2017.

[13] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[14] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1335-1344.

[15] Roth P M, Hirzer M, Koestinger M, et al. Mahalanobis distance learning for person re-identification[M]//Person Re-Identification. Springer London, 2014: 247-267.

[16] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[C]//AAAI. 2017: 4278-4284.

[17] Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[J]. arXiv preprint arXiv:1704.01719, 2017.

[18] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[19] Varior R R, Shuai B, Lu J, et al. A siamese long short-term memory architecture for human re-identification[C]//European Conference on Computer Vision. Springer International Publishing, 2016: 135-153.

[20] Zhong Z, Zheng L, Cao D, et al. Re-ranking Person Re-identification with k-reciprocal Encoding[J]. arXiv preprint arXiv:1701.08398, 2017.

[21] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116-1124.

[22] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 248-255.

[23] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[24] Sun Y, Zheng L, Deng W, et al. SVDNet for Pedestrian Retrieval[J]. arXiv preprint arXiv:1703.05693, 2017.

[25] Koestinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2288-2295.

[26] Zhao R, Ouyang W, Wang X. Learning mid-level filters for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 144-151.

[27] Xiao T, Li H, Ouyang W, et al. Learning deep feature representations with domain guided dropout for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1249-1258.