2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)

Person Re-identification with Deep Features and Transfer Learning

Shengke Wang[1], *Shan Wu*[1], *Lianghua Duan*[1],*Changyin Yu*[1], *Yujuan Sun*[2], *Junyu Dong*[1]
[1]Department of Computer Science and Technology, Ocean University of China,
[2]School of Information and Electrical Engineering, Ludong University,
neverme@ouc.edu.cn , 614332478@qq.com , 1160050472@qq.com, 974657627@qq.com, syj_anne@163.com,
dongjunyu@ouc.edu.cn

***Abstract*—**Person re-identification is an important technique towards automatic search of a person's presence in a surveillance video. Two fundamental problems are critical for person re-identification:feature representation and metric learning. At present, there are many methods in the study of person re-identification, which has achieved remarkable results. Due to the difference of the data distribution in different scenarios, the performance of the person re-identification in the new scene is significantly decreased. In order to avoid the tedious manual annotation, and to make full use of the original detector and labeled samples, the research of person re-identification based on transfer learning has received more and more attention. Existing approaches adopt a fixed metric for matching all the subjects. In this work, we propose a Feature Net (FN) architecture with Convolution Neural Networks (CNNs) to learn the pedestrian feature, reserved more useful information. And use Cosine distance to measure the each image pair's similarity directly which is more efficient but uncomplicated than others. Our method can be applied to different scenarios and improved the recognition performance. Experiments on the challenging datasets show the effectiveness of our methods, especially on cuhk03 dataset, we achieve the state-of-the-art result.

Keywords—re-identification;transfer learning;deeping learning; distance measure;

I. INTRODUCTION

Person re-identification refers to the multi camera surveillance system in the non-overlapping field of vision, which matches the pedestrian target in the vision of different cameras. It's been a lot of attention over the years because of its important applications in video surveillance, such as cross-camera tracking, multi-camera behavior analysis and pedestrian search. In the course of the study, the Person re-identification [1] is faced with the challenge from the viewpoint, illumination and occlusion, which makes the task more difficult. After continuous research, many methods have achieved good results, which can be applied to real life. Person re-identification technology is the first feature extraction, and then use the metric learning method for feature matching. For feature selection, several effective approaches have been proposed, for example,HOG [2], ELF[3],LBP[4] and SIFT[5].

Besides, how to learn a robust distance or similarity metric has become another topic of concern for person re-identification. The main purpose is to find an accurate distance or similarity calculation to determine which of the two images is the same person. Most methods first adapt Principle Component Analysis to reduce the dimension, then metric learning is performed on the PCA subspace. These methods are to learn the distance metrics to match the visual features of image regions observed in different camera views, including Cross-view Quadratic Discriminant Analysis (XQDA) [6], Euclidean distance metric learning, and Cosine distance metric learning. Besides most of these method assume that the training and test samples are captured in similar scenarios so that their distributions are assumed to be the same. This assumption is not true in many real visual recognition applications.

In order to solve the above problems ,in this paper we propose a Feature Net (FN) architecture with Convolution Neural Networks (CNNs) to learn the pedestrian feature for cross-dataset visual recognition. And then, we take Cosine distance as our similarity measures to calculate images' distance directly. The feature processed by FN is 64 dimension, reserved more useful information but less noise.With omitting the process of metric learning, we reduce the process time while improved the accuracy.

II. RELATED WORK

A. Person Re-identification

With the construction and improvement of the large-scale video camera surveillance network related to urban public safety, the re-identification and search of the specific pedestrian targets in the video surveillance network is becoming more and more important issue. Person re-identification is the task of matching two pedestrian images from different viewpoints, which extracting reliable, robust, computable characteristics from pedestrian original image, then build the descriptive and discriminative model of the pedestrian's visual patterns.

Studying on person re-identification usually focuses on finding an improved set of features.Feature representation methods aim to seek discriminative descriptors which are robust to variations of viewpoint, pose, and illumination in pedestrian images captured across different cameras [7].Features that have been used include variations on color histograms[8], local binary patterns[9], Gabor features[10]. Many excellent methods have been proposed. For example, Liao et al. [6] improved the KISSME method by learning a discriminant low dimensional subspace based on the LOMO

978-1-5386-3221-5/17 $31.00 © 2017 IEEE
DOI 10.1109/CSE-EUC.2017.136

features, and maximizes the occurrence to make a stable representation against viewpoint changes. Farenzena et al. [11] developed a symmetry driven accumulation of local feature for appearance modeling of human body images. Cheng et al.[12]utilized the Pictorial Structures where part-based color information and color displacement were considered for person re-identification.

There are also many metric learning algorithms [13] designed for person re-identification. They are designed to learn a distance metric to reduce the distance of the matched images, and enlarge the distance of the mismatched images. Li at al. [14] proposed the learning of Locally-Adaptive Decision Functions (LADF) for person verification, which can be considered as a joint model of a distance metric and introducing an adaptive threshold into Mahalanobis distance. Prosser et al. [15] viewed the person re-identification problem as a ranking problem, and used the RankSVM to learn a subspace. Weinberger et al. [16] proposed a large margin nearest neighbor (LMNN) model, where the distance metric is learned for large margin nearest neighbor classification and to separate the matched neighbors from the mismatched ones by a large margin.

B. *Transfer learning*

When the distribution of the training data from the source domain is different from the distribution of the training data for the target domain, transfer learning is intended to solve the problem. At present, transfer learning has been widely used in the fields of Natural Language Processing and pattern recognition.It can be mainly categorized into two classes: instance-based [17] and feature-based [18]. For the first class, different weights are learned to rank the training samples in the source domain for better learning in the target domain. For the second class, a common feature space is usually learned which can transfer the information learned from the source domain to the target domain. In recent years, several representative transfer learning methods [19] have been proposed.Cross-dataset transfer learning [20] has been employed for person re-identification, and it is desirable that labelled data from other camera datasets can provide transferable identity discriminative information for the unlabelled target dataset.

III. TRANSFERING METRIC LEARNING

A. *Deep Transfer Metric Learning*

Hu et al. [20] proposed Deep Transfer Metric Learning which learns a set of hierarchical non-linear transformations. It is able to transfer a view-invariant representation from the labeled source domain to the unlabeled target domain. This requires the inter-class variations are maximized and the intra-class variations are minimized.

For each pair of samples x_i and x_j , they can be finally represented as $f^{(m)}(x_i)$ and $f^{(m)}(x_j)$ at the mth layer of our designed network, and their distance metric can be measured by computing the squared Euclidean distance between the representations$f^{(m)}(x_i)$and$f^{(m)}(x_j)$ at the mth layer:

$$d^2_{f(m)}(x_i, x_j) = ||f^{(m)}(x_i) - f^{(m)}(x_j)||_2^2 . \quad (1)$$

Following the graph embedding framework, we enforce the marginal fisher analysis criterion on the output of all the training samples at the top layer and formulate a strongly-supervised deep metric learning method as:

$$\min_{f^{(M)}} J = S_c^{(M)} - \alpha S_b^{(M)} + \gamma \sum_{m=1}^{M} (||W^{(m)}||_F^2 + ||b^{(m)}||_2^2) \quad (2)$$

Where α(α>0)is a free parameter which balances the important between intra-class compactness and interclass separability; $||Z||_F$ denotes the Frobenius norm of the matrix Z;γ(γ>0)is a tunable positive regularization parameter; $S_c^{(m)}$and $S_c^{(m)}$ define the intra-class compactness and the interclass separability, which are defined as follows:

$$S_c^{(m)} = \frac{1}{Nk_1} \sum_{i=1}^{N} \sum_{j=1}^{N} P_{ij}\, d^2_{f^{(m)}}(x_i, x_j), \quad (3)$$

$$S_b^{(m)} = \frac{1}{Nk_2} \sum_{i=1}^{N} \sum_{j=1}^{N} Q_{ij}\, d^2_{f^{(m)}}(x_i, x_j), \quad (4)$$

where P_{ij} is set as one if x_j is one of k_1-intraclass nearest neighbors of x_i, and zero otherwise; and $Q_{ij}$is set as one if x_j is one of k_2-interclass nearest neighbors of x_i, and zero otherwise.

Given target domain data $X_t$and source domain data X_s, their probability distributions are usually different in the original feature space when they are captured from different datasets. To reduce the distribution difference, it is desirable to make the probability distribution of the source domain and that of the target domain be as close as possible in the transformed space. To achieve this, they apply the Maximum Mean Discrepancy (MMD) criterion to measure their distribution difference at the mth layer, which is defined as follows:

$$D_{ts}^{(m)}(x_t, x_s) = || \frac{1}{N_t} \sum_{i=1}^{N_t} f^{(m)}(x_{ti}) - \frac{1}{N_s} \sum_{i=1}^{N_s} f^{(m)}(x_{si}) ||_2^2 \quad (5)$$

By combining (2) and (5), they formulate DTML as the following optimization problem:

$$\min_{f^{(M)}} J = S_c^{(M)} - \alpha S_b^{(M)} + \beta D_{ts}^{(m)}(x_t, x_s) + \gamma \sum_{m=1}^{M} (||W^{(m)}||_F^2 + ||b^{(m)}||_2^2) \quad (6)$$

Where β(β≥ 0) is a regularization parameter.

B. *Our Method*

In this paper, we propose a simple method to solve the person re-identification problem. In our work, we use a very deep Convolution Neural Network (according to the Microsoft research, the CNN's result the deeper the better) structure to extract the pedestrian feature and then apply it to the Cosine Distance method directly. Fig. 1 illustrates our architecture. First, we use CNN part to get the better feature representation for pedestrian images by training the classification net, then we utilize the "fc6" layer's output as the images feature and put it in Distance part to calculate the similarity of image pairs.

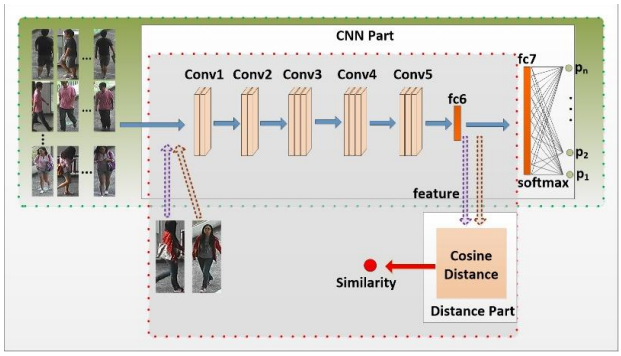

Fig. 1. The overview of our method. For the person re-identification issue, we first train a classification CNN for feature extraction (green box), then calculate images' similarity by a Distance part (red box).

In the section of feature extract, we design our Feature Net (FN) modeled on the VGG16 Net-work [21] which have been achieved state-of-the-art performance on ImageNet dataset. Detailed structures are listed in Table. 1 ("N" denote that the numbers of dataset's train set, "conv+BN+ReLU" means that it consist of a convolution layer, a Batch Normalization layer and a ReLU layer, "FC" means a full connected layer). We use the Batch Normalization (BN) strategy accelerate the convergence process before the ReLU layer referenced in [22].In order to improve the robust of our CNN framework, we use Dropout strategy [23], which is wide used in the CNN's train step, dropout some neurons before processing the last full connected layer.

Table 1. THE STRUCTURE DETAILS OF CNN STRUCTURE

name		Kernel size / stride /pad	output size
input			3 × 144 × 56
conv1	(conv+BN+ReLU)*2	3/1/1	64 × 144 × 56
	Maxpool1	2/2	64 × 72 × 28
conv2	(conv+BN+ReLU)*2	3/1/1	128 × 72 × 28
	Maxpool2	2/2	128 × 36 × 14
conv3	(conv+BN+ReLU)*3	3/1/1	256 × 36 × 14
	Maxpool3	2/2	256 × 18 × 7
conv4	(conv+BN+ReLU)*3	3/1/1	512 × 18 × 7
	Maxpool4	2/2	512 × 9 × 4
conv5	(conv+BN+ReLU)*3	3/1/1	512 × 9 × 4
	Maxpool5	2/2	512 × 5 × 2
fc6	FC+ReLU+Dropout		64
fc7	FC+ReLU		N

In the section of distance metric learning, we use Cosine distance as similarity measure. $A = (\alpha_1, \alpha_2, \cdots, \alpha_{64}), B = (b_1, b_2, \cdots, b_{64})$represent the two features extracted by our FN from two images, higher values indicate a higher similarity of two images, the similarity of two images can be denote as:

$$similarity = cos(\theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{64} A_i B_i}{\sqrt{\sum_{i=1}^{64} A_i^2}\sqrt{\sum_{i=1}^{64} B_i^2}} \quad (7)$$

IV. EXPERIMENTS

We implemented our Feature Network (FN) architecture using the Caffe deep learning framework. Network training 50k iteration in CUHK03 [24] converge in roughly 10 hours on NVIDIA Titan x, finetune on other dataset expend 3-5 hours for 1.5k iteration. We report a comprehensive evaluation of our method by comparing it to the state-of-the-art approach on various data set (CUHK03 [24], CUHK01 [25], VIPeR [26]).

A. Experimental details

When operating the Feature Network (FN), we employ mini-batch stochastic gradient descent (SGD) for faster back propagation and smoother convergence. In each iteration of training phase, the mini-batch is 50 images, learning rate =0.01, and decreased by every 20,000 iteration. We train CUHK03 dataset firstly due to it's the largest dataset and then finetune it on other datasets. The dataset and protocols are followed at Table 2 ("#" means "the number of", "ID" means "the identity of person", "FN"means "our Feature Network", "Tr" means "train", "Val" means "validation", "Pro" and "Gal" means the probe and gallery dataset for test). With the procedure of FN accomplished, we forward the probe and gallery data and save the fc6 layer's output as the images' feature.

Table 2.THE DETAILS OF DATASET

Dataset	#ID	#FN			#Test	
		#FN ID	*#Tr images*	*#Val images*	*#Pro ID*	*#Gal ID*
CUHK03	1467	1367	21009	5252	100	100
CUHK01	971	486	1552	388	485	485
VIPeR	632	316	506	126	316	316

B. Experimental results

The CUHK03 [24] data set has more than 14,000 images of 1467 subjects, captured by five different pairs of surveillance cameras. Each identity has 10 images approximately. Following the protocol, we randomly partition 1467 pedestrians into non-overlap FN set, Test sets. We use 21009 train images and 5252 validation images to train our classification Network, 50k iteration late, the accuracy of classification can achieved 99.8%. Based on the learned model, we evaluate the test set to illustrate the efficiency of proposed method. We compare our approach against traditional method which gained the previously best performance on CUHK03 with rank-1 matching up to75.3%. The CMC curves and the rank-1 identification rates are shown in Fig 2 (left). From picture, we can see our method is superiority over others both traditional and new method by CNN. We achieves the state-of-the-art result with rank-1 75.4%.

The CUHK01 [25] data set contains 971 subjects captured from two camera views in a campus environment. We set the number of individuals in the train split to 486 and test split to 485.We initialize the Feature Network by the model pre-trained on CUHK03 dataset, and then finetune the FN on CUHK01. After1.5k iteration, we extract the features of probe and gallery dataset using the trained model, followed by the Cosine Distance, we can get the similarity of probe and gallery images. The CMC curve and rank-1 identity rate is shown in Fig. 2 (right).

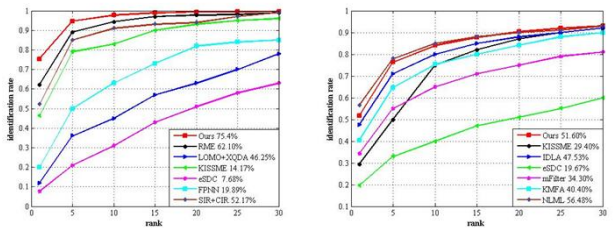

Fig. 2. CMC curves and rank-1 identification rates on CUHK03 dataset (left) and CUHK01 (right) dataset. Our method achieve the state-of-the-art result on CUHK03 dataset.

The VIPeR [26] data set contains 632 pedestrian pairs in two views, with only one image per person in each view, is especially challenging for the reasons of viewpoint and low-solutions. Based on the protocol, we finetune the dataset using the CUNHK03 dataset's model, after the 1.5K iteration, the result is drawn in Fig. 3. Our method win the best while has a wide gap with the sate-of-the-art result 47.8%.

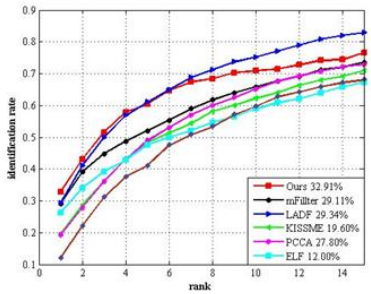

Fig. 3. Cmc curve on VIPeR dataset

V. CONCLUSION

In this paper, on the basis of deep transfer learning, we apply the transfer learning to the study of Person re-identification, present an effective way of feature extraction for person re-identification, and use the CNN that can collect the available information automatically with the BP strategy, called Feature-Net (FN). The feature extracted by FN is low-dimension but valid for calculate the similarity of two images. What's more, we employ the simple but useful Cosine Distance as the measure method directly saving the time while performing excellently. Experiment on three challenging and common person re-identification datasets, we beat the most method individually, demonstrates the robust of our method, especially on CUHK03 we obtain the state-of-the-art result with rank-1 accuracy of 75.4% .

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of China (NSFC) Grants 61301241, 61403353, 61501417 and 61271405; Natural Science Foundation of Shandong (ZR2015FQ011; ZR2014FQ023); China Postdoctoral Science Foundation funded project (2016M590659); Qingdao Postdoctoral Science Foundation funded project (861605040008);The Fundamental Research Funds for the Central Universities (201511008, 30020084851).

REFERENCES

[1] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In ICCV, IEEE Conference on,2013, pp. 2528–2535.

[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection,in:CVPR,2005.pp. 886-893.

[3] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In ECCV, 2008.

[4] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24 (2002) pp.971-987.

[5] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision, 60 (2004) 91-110.

[6] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by Local Maximal Occurrence representation and metric learning, In Computer Vision and Pattern Recognition ,IEEE Conference on, 2015.

[7] L. Bazzani, M. Cristani, A. Perina. Multipleshot person re-identification by chromatic and epitomic analyses. PRL,2012,pp. 898–903.

[8] N. Martinel, C. Micheloni, and G. Feresti. Saliency weighted features for person re-identification. In ECCV Workshop on Visual Surveillance and Re-identification, 2014.

[9] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In ECCV, 2014.

[10] W. Li and X. Wang. Locally aligned feature transforms across views. In CVPR, 2013.

[11] Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features. Computer Vision and Pattern Recognition. DBLP, 2010.pp.2360-2367.

[12] Dong S C, Cristani M, Stoppa M, et al. Custom Pictorial Structures for Re-identification. Bmvc, 2011:68.1-68.11.

[13] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon.Information theoretic metric learning. In ICML, 2007,pp. 209-216.

[14] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R.Smith. Learning locally-adaptive decision functions for person verification. In IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[15] Engel C, Baumgartner P, Holzmann M, et al. Person Re-Identification by Support Vector Ranking. British Machine Vision Conference, BMVC 2010,pp.1-11.

[16] Weinberger K Q, Saul L K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. NIPS. 2006,pp.207-244.

[17] W. Dai, Q. Yang, G. Xue, and Y. Yu. Boosting for transfer learning. In ICML, 2007,pp.193–200.

[18] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. JMLR, 2005,6:1817–1853.

[19] Y. Zhang and D. Yeung. Transfer metric learning by learning task relationships. In KDD, 2010,pp.1199–1208.

[20] Hu J, Lu J, Tan Y P. Deep transfer metric learning. Computer Vision and Pattern Recognition. IEEE, 2015,pp.325-333.

[21] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science, 2014.

[22] Wu S, Chen Y C, Li X, et al. An enhanced deep feature representation for person re-identification. IEEE Winter Conference on Applications of Computer Vision. IEEE, 2016,pp.1-8.

[23] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. Computer Science, 2012, 3(4):pp. 212-223.

[24] Li W, Zhao R, Xiao T, et al. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In Computer Vision and Pattern Recognition ,IEEE Conference on, 2014,pp.152-159.

[25] Li W, Wang X. Locally Aligned Feature Transforms across Views. Computer Vision and Pattern Recognition. IEEE, 2013,pp.3594-3601.

[26] Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking. Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS). 2007, 3(5).