Cross Dataset Person Re-identification

Yang Hu, Dong Yi, Shengcai Liao, Zhen Lei, Stan Z. Li*

Center for Biometrics and Security Research
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences (CASIA)
95 Zhongguancun East Road, 100190, Beijing, China
{yhu, dong.yi, scliao, zlei, szli}@nlpr.ia.ac.cn

Abstract. Until now, most existing researches on person re-identification aim at improving the recognition rate on single dataset setting. The training data and testing data of these methods are form the same source. Although they have obtained high recognition rate in experiments, they usually perform poorly in practical applications. In this paper, we focus on the cross dataset person re-identification which make more sense in the real world. We present a deep learning framework based on convolutional neural networks to learn the person representation instead of existing hand-crafted features, and cosine metric is used to calculate the similarity. Three different datasets Shinpuhkan2014dataset, CUHK and CASPR are chosen as the training sets, we evaluate the performances of the learned person representations on VIPeR. For the training set Shinpuhkan2014dataset, we also evaluate the performances on PRID and iLIDS. Experiments show that our method outperforms the existing cross dataset methods significantly and even approaches the performances of some methods in single dataset setting.

1 Introduction

Person re-identification has attracted more and more attention in recent years. It aims to recognize individuals through person images taken from two or more non-overlapping camera views. As an important and basic component in surveillance system, person re-identification is closely related to many other applications, such as cross-camera tracking, behavior analysis, object retrieval and so on. However, the person re-identification problem is a very challenging task. The person images of existing datasets are captured from surveillance cameras which usually set to work in the wide-angle mode to cover a wider area, therefore the resolution of person images is very low even using the high-def cameras. Moreover, changes in illumination, viewpoint, background, pose, camera parameter and occlusion under different camera views make the person re-identification a difficult problem: (1) lack of samples to generate the true distributions of various classes; (2) the distributions of the intra classes and inter classes are unstable

[*] Stan Z. Li is the corresponding author

due to the diversities and ambiguities of the samples; (3) above all, the samples of person re-identification datasets are inseparable.

In the past few years, researchers have done a lot meaningful work to advance the development of person re-identification. Many difficulties of the task have been solved to some extent by discriminative hand-crafted features and metric learning methods. However, the cross dataset setting is different, which has been ignored by most existing methods. The testing sets and training sets are not the same, we don't know the testing data as well as the acquisition conditions, many important factors such as illumination condition, viewpoint and occlusion are totally unconstrained therefore make the task more challenging.

Most existing person re-identification methods conduct their experiments on several public datasets. The most common way is spitting the dataset into two parts, one part for training, the other part for testing. In this setting, the training data and testing data are from the same source. However in real applications, it is very difficult to get a training set which has a similar scenario of the testing set. Therefore unchanged view information and similar data construction make most learning based methods have bad generalization and easy to over-fitting. For feature design based methods, it is also difficult to ensure that the designed features to be effective for new coming data. In a work, most existing methods do not perform well.

Since the performances of single dataset person re-identification methods are improving rapidly, researchers should pay more attention to the cross dataset person re-identification task which make more sense to the real applications. In this paper, we try to solve the cross dataset person re-identification problem by the deep learning (DL) framework which has got huge successes in speech recogintion and vision. DL is especially suitable for dealing with large training sets, most existing public person re-identification datasets are small, both in the number of subjects and the number of images per subject. However, with the development of person re-identification, more and more datasets have emerged. The scales of these datasets are getting larger although are still not comparable to other fields [1, 2]. Three datasets Shinpuhkan2014dataset[3], CUHK02[4, 5] and CASPR are selected as the training sets, experiments will show the performances of the learned person representations as well as the impact of different training set structures. Compared to the standard feedforward neural networks, convolutional neural networks (CNNs) have much fewer connections and parameters and easier to train. Therefore, we choose CNNs to learn a generic representation of person images.

In this paper, we make the following contributions: (1) We present a CNN framework to learn effective features for person re-identification which has not been paid much attention before; (2) As a different and important aspect of person re-identification, the cross dataset person re-identification advocated by this paper is very meaningful for real applications; (3) The person representations obtained by the proposed method have good performances and generalize well to many datasets.

2 Related Work

The recognition rates of person re-identification have increased a lot over the lase several years. Among these methods, metric learning (ML) approaches have played very important roles[6–13]. Weinberger et al.[6] proposed the LMNN method to learn a Mahanalobis distance metric for k-nearest neighbor (kNN) classification by semidefinite programming. Subsequently, in [7], a method similar to LMNN called Large Margin Nearest Neighbor with Rejection (LMNN-R) was proposed and achieved significant improvement. Davis et al.[8] presented an approach called LTML to formulate the learning of Mahalanobis distance function the problem as that of minimizing the differential relative entropy between two multivariate Gaussians under constraints on the distance function. Zheng et al.[9] formulated person re-identification as a distance learning problem, which aimed to learn the optimal distance that can maximize the probability that a pair of true match having a smaller distance than a wrong match pair. Koestinger et al.[10] proposed the KISSME method to learn a distance metric from equivalence constraints based on a statistical inference perspective. Li at al.[11] proposed the Locally-Adaptive Decision Functions (LADF) method to learn a decision function for person verification that can be viewed as a joint model of a distance metric and a locally adaptive thresholding rule. However, most of these methods have shown to be sensitive to parameters selecting and very easily over-fitting, the performances are not satisfactory in real applications.

Another type of person re-identification methods try to tackle the problem by seeking feature representations which are both distinctive and stable for describing the appearance[14–17]. Farenzena et al.[18] proposed the Symmetry-Driven Accumulation of Local Features (SDALF) method, multiple features were combined considering the symmetry and asymmetry property in pedestrian images to handle view variation. Malocal et al.[19] turned the local descriptors into Fisher Vector to produce a global representation of the image. Cheng et al.[20] utilized Pictorial Structures for person re-identification. Color information and color displacement within the whole body were extracted per-part, and the extracted descriptors were then used in a matching step. Salience were gradually applied in person re-identification as well[21–23]. However, most of hand-crafted features are not distinctive and stable enough and may lose efficacy due to various factors, such as illumination, viewpoint and occlusion, especially when the data sources have changed.

There are also other person re-identification methods. Gray and Tao [24] proposed to use AdaBoost to select good features out of a set of color and texture features. Prosser et al.[25] formulated the person re-identification problem as a ranking problem and applied the Ensemble RankSVM to learn a subspace where the potential true match get the highest rank. In [26], visual features of an image pair from different views are first locally aligned by being projected to a common feature space and then matched with softly assigned metrics which are locally optimized. Liu et al.[27] allowed users to quickly refine their search, which achieved significant improvement. In conclusion, most existing person re-identification methods pay their attentions on single dataset setting, they all contribute a lot to promote the development of this field, but the performances in real applications are still not good enough.

Fortunately, some researchers have noticed the cross dataset setting and have done some meaningful work. Ma et al.[28] proposed a Domain Transfer Ranked Support Vector Machines (DTRSVM) method for re-identification under target domain cameras which utilized the image pairs of the source domain as well as the unmatched (negative) image pairs of the target domain. Although this method cannot be totally considered as cross dataset person re-identification (used the information of the target domain), it inspired us a lot and got impressive results.

Yi at al.[29] proposed a method called Deep Metric Learning (DML) which learn the metric by a "siamese" deep neural network. The network had a symmetry structure with two sub-networks which were connected by a cosine layer. In this paper, the cross dataset person re-identification experiment were conducted, the author utilized the CUHK Campus dataset as the training set and the ViPeR dataset for testing. Big improvement has been made by DML compared with the DTRSVM on VIPeR. Moreover, DML is the first to conduct experiments on cross dataset person re-identification, although further research is needed, it has promoted the development of cross dataset person re-identification a lot.

Due to the obvious superiorities on large amount of training data and the development of computation resources such as GPU, more and more researchers are carrying on researches on CNN. Therefore, CNN has achieved great success in many fields of computer vision, for instance, DeepFace[1] and [2] have exhibited impressive results on face recognition and image classification. Inspired by DeepFace, we will address the cross dataset person re-identification problem by learning the person representation. In the following sections, the implementation details will be described, the comparison experiments and the discussions will be reported as well.

3 Person Representation

In recent years, as more data has become available, learning-based methods have started to outperform hand-crafted features, because they can discover and optimize features for the specific task. For cross dataset person re-identification, the influence of resolution, illumination and pose changes are much greater. Moreover, the change of data sources may make the well designed features perform well on one dataset but worse on another. Here, we address the problem by learning the person representation through deep neutral network.

3.1 The Architecture of CNN

The flowchart of our method is shown in Figure 1. Given a person image, it is first normalized to 48 by 128 pixels, and then divided into three overlapped parts of size 48 by 48 pixels respectively. For each person part, we train a CNN and extract features of this part. Finally, we concatenate the three learned features to get the final person representation.

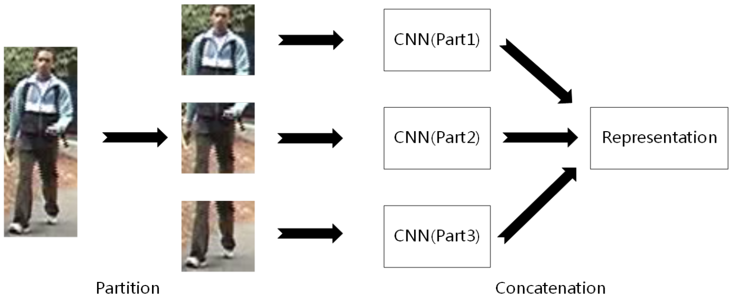

Fig. 1. The flowchart of the proposed method. Learn the representation for each part independently and concatenate them to get the final representation.

As the way used for handwritten digit recognition, we train our CNN in a multi-class classification manner. The architecture of our neural network is summarized in Figure 2. The input of our network is the raw 3-channels (RGB) person image part of size 48 by 48 pixels and send to a convolutional layer (C1). The number of filters in C1 is 32 and the size of the filters is 5 × 5. Then the 32 feature maps are fed to a max-pooling layer (S2) which takes the max value over 3 × 3 spatial neighborhoods with a stride of 2 for each channel. Followed by the S2 is another convolutional layer (C3) with 32 filters and the size of the filters is 5 × 5. The first three layers can extract low-level features such as simple edges and texture. Max-pooling layers play an important role in dealing with the local rotations and transformations, we apply a max-pooling layer after each convolutional layer to make the network more robust. In some other CNN framework, the max-pooling layer is only applied in the first convolutional layer to avoid losing information about the precise position of detailed structure and micro-textures. However, in our task, persons have a variety of changes in pose and the range of variation is very wide. Moreover, no alignment is applied to the unconstrained persons. Therefore, we apply a max-pooling layer after each convolutional layer to make sure the learned representations more general and robust.

The structure of the subsequent layers are just like the former layers. There are two convolutional layers C3 and C5 with 32 filters and 64 filters respectively, the size of the filters is 5 × 5. Here, we will introduce why we choose this structure. In many other work, local layers are used to ensure that every location in the feature map learns a different set of filters. There are two reasons why local layers are applied: (1) In an aligned image, different regions may have different local statistics and discrimination abilities, the spatial stationarity assumption of convolution cannot hold; (2) Large labeled datasets are available which can afford large locally connected layers such as in the face recognition field. Therefore, the local connected layers are reasonable. However, in person re-identification field, person alignment is a very difficult problem and there is no existing effective algorithm to use. On the other hand, the scales of datasets are much smaller which are not enough to learn the huge number of parameters of local connected layers. Therefore, we choose convolutional layers instead of local layers which sharing the weight for the cross dataset person re-identification task. It may sacrifice some discriminations but is a safe solution to make the learned representation more general and robust.

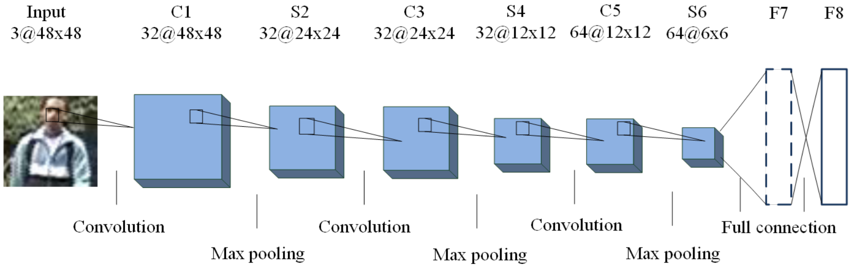

Fig. 2. The structure of the CNN used in our method.

The top two layers (F7 and F8) are fully connected. The first fully connected layer (F7) is optional, the output of the last fully connected layer (F8) is fed to a K-way softmax which produces a distribution over K class labels. If the training set has a small number of subjects, F7 is removed and the output of F8 (after softmax) is used as the feature vector which has a dimension of the number of subjects. If the training set has a relatively large number of subjects, F7 is needed for dimension reduction to avoid over-fitting, and the output of F7 in the network will be used as feature vector. More details will be discussed in the experiments.

3.2 Training and Verification Metric

The training aims to maximize the probability of the correct person identity by maximizing the multinomial logistic regression objective. We achieve this by minimizing the cross-entropy loss for each training sample. If k is the index of the true label for a given input, the loss is: $L = -logp_k$. The loss is minimized by stochastic gradient descent with the size of batch 128. The gradients are computed by standard backpropagation of the error[30]. Fast learning has a great influence on the performance of models trained on large datasets. Therefore, we choose the ReLU[31] to be our activation function: max(0,x) for each layer.

After we get the person representation, we can apply a simple metric, such as the cosine distance, Euclidean distance and Battacharrya distance, to calculate the similarity of two input instances. In this paper, we pay attention to learn general and robust person representation for cross dataset person re-identification, the classifier design is not the focus. Therefore, for the proposed method, we use the cosine distance. Experiments can show that even with simple metric, our learned person representation can perform well.

4 Experiments

In this section, we first introduce the datasets used in our experiments and then the training details, after that we present the comparison with the state-of-the-art, some findings and discussions of the experiments are presented as well.

4.1 Datasets

Three datasets Shinpuhkan2014dataset, CUHK02 and CASPR are chosen as the training sets. We get three representations from the CNN models trained by these three datasets respectively. Then, we evaluate the performances of the representations on VIPeR[32] compared with two other existing cross dataset person re-identification methods DTRSVM[28] and DML[29]. After that, we fix the training set to Shinpuhkan2014dataset, and evaluate performance of the learned representation on PRID[33] and i-LIDS[9].

Shinpuhkan2014dataset consists of more than 22,000 images of 24 people which are captured by 16 cameras installed in a shopping mall "Shinpuh-kan". All images are manually cropped and resized to 48 × 128 pixels, grouped into tracklets and added annotation. The number of tracklets of each person is 86. This dataset contains multiple tracklets in different directions for each person within a camera. The greatest advantage of this dataset is that all the persons have appeared in 16 cameras. Some image samples can be seen in Figure 3.

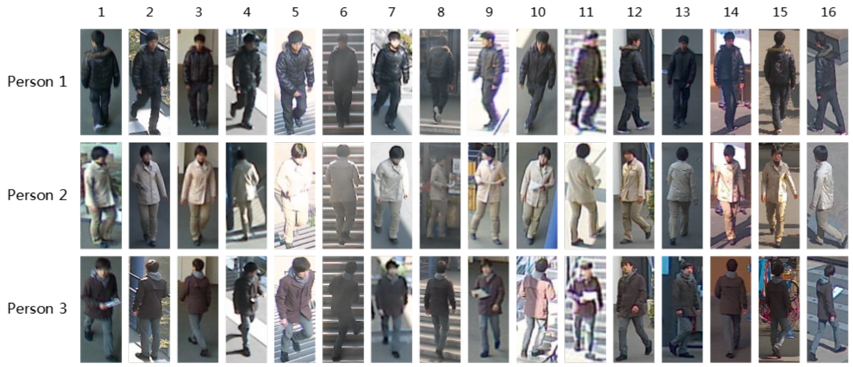

Fig. 3. Some image samples of three persons selected from the Shinpuhkan2014dataset, images from the same column indicate that they are from the same camera view. We can see that for each person, the appearances are different in 16 cameras and there are many kinds of changes.

CASPR is a person re-identification dataset collected by ourselves. We capture six videos in a research institution and segment 7414 images of 200 persons from these videos. Note that each of the 200 person has at least 2 associated camera views, some of them have appeared in 5 camera views. All segmented images are scaled to 48 × 128 pixels. Each subject has at least 7 images and at most 93 images. Figure 4 shows some example pairs of images from the collected database. It can be observed that there is a large variation in the observed color, and there are also lighting changes and viewpoint changes that challenge the matching of persons across cameras. We are expanding this dataset and it will be released in the near future.

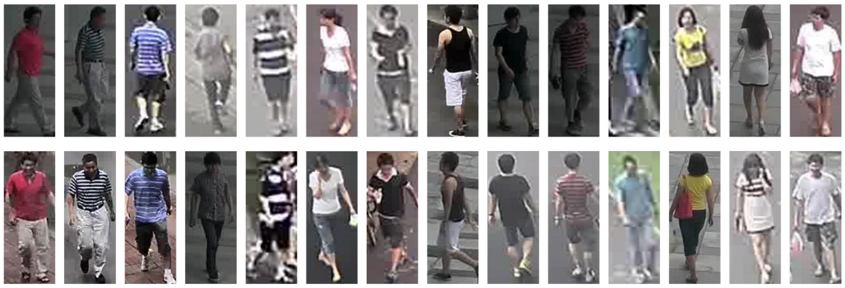

Fig. 4. Example pairs of images from the collected database. Images in the same column come from the same person.

CUHK02 contains 1,816 persons captured from 5 pairs of cameras (P1-P5,ten camera views). They have 971, 306, 107, 193 and 239 persons respectively. Each identity has two samples per camera view. It has the largest number of subjects so far. Samples from this dataset can be found in Figure 5.

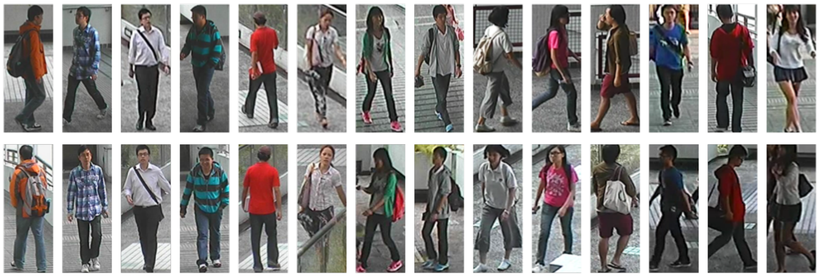

Fig. 5. Example pairs of images from the CUHK02 database. CUHK02 has five pairs of camera views denoted with P1-P5. Here, at lest two exemplar persons are shown for each pair of views. Images in the same column represent the same person.

VIPeR The Viewpoint Invariant Pedestrian Recognition database is one of the earliest single-shot datasets, and it is the most widely used benchmark so far in person re-identification field. It contains 632 pairs of pedestrians and images in VIPeR suffer greatly from illumination and viewpoint changes, making it a very challenging dataset.

4.2 Training on the Datasets

Firstly, we train the deep neural network on the Shinpuhkan2014dataset by the Tesla GPU. As the number of subjects is very small (24 persons), we take out the first fully connected layer (F7) and consider the output of the last fully connected layer (F8, after a 24-way softmax) to be our feature vector. We initialize the weights in each layer from a normal distribution with mean zero and standard deviation 0.0001, and the biases are set to 0.5. The weight decay of the F8 is set to 3. The stride is set to 1 and the padding is set to 2. During the training, we flip the images to double the number of training samples, and use the training set to test every 5 epoches. On our experience, as the number of classes (subjects) in this network is small which makes the cost of the test set drop easily, we set the epoch to 100 for this dataset with a small learning rate (0.001 for weight and 0.002 for bias).

For the CUHK02 dataset, the first fully connected layer (F7) is needed and the output of F7 is used as our feature vector. We set the dimension of F7 to 24 and we can get a 24-dimension feature vector for each part of a input image (see the division of a person in Fig.1). Most of the parameters are same with the parameters when we train the model on Shinpuhkan2014dataset except the number of epoches. As the numbers of subjects are bigger and the numbers of images per subject are smaller, the error rate is much more difficult to drop. Therefore, we set the epoch to 250 for this dataset.

For the CASPR dataset, the number of subjects and the number of images per subject is the middle of that of the Shinpuhkan2014dataset and the CUHK02. The error rate is also hard to drop. We apply almost the same structure and parameters of CUHK02 except the epoches. A more efficient way is to apply a two-step strategy to train the model which is also feasible for the CUHK02 dataset. The first step, we set a relatively large learning rate (0.01 for weight and 0.02 for bias) to make the error rate drop, when the it start to drop, we change the learning rate to a smaller value (0.001 for weight and 0.002 for bias) to make the error rate drop slowly and smoothly. The number of epoch is 60 and 100 for step one and step two respectively. This strategy can save the training time as well as preventing the oscillation of error rate which make it hard to drop to a more optimal point.

4.3 Results on VIPeR

For each of the three training sets, we show the performances of the learned person representations on VIPeR. We split VIPeR into testing set with 316 subjects randomly, and repeat the process 11 times. The first split is used for parameter tuning, the other 10 splits are used for reporting the results.

Figure 6, Figure 7 and Figure 8 show the rank curves of the three parts and their fusion. The recognition rates are summarized in Table 1 as well as the comparison with DTRSVM and DML. The most difficult point is the difference in data distribution of different datasets, the model learned on one dataset probably lose efficacy on new data. From the results we can see that the cross dataset

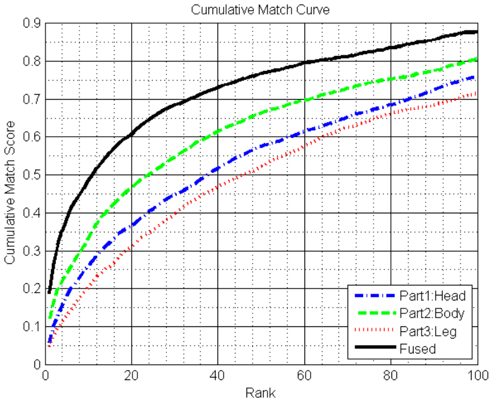

Fig. 6. The CMC curves on the VIPeR dataset, the training set is Shinpuhkan2014dataset. Performances of each body parts and the fusion are shown.

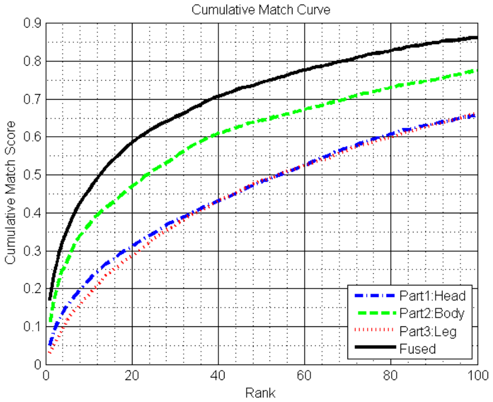

Fig. 7. The CMC curves on the VIPeR dataset, the training set is CUHK02.

evaluation accuracies of person re-identification are currently very low. For our method, although the training sets are different, it develops a general feature representation that can be directly applied in different scenarios, which is very important for practical applications. Our method gets very impressive results and even approaches the performance of some methods in single database setting, such as ELF [24] and PRDC [9]. Shinpuhkan2014dataset has large variations such as viewpoint, background, illumination, pose and deformation due to the 16 different camera views, the number of images is large as well. Although the number of subjects is not large enough, this dataset is very suitable for learning the person representation.

We also compared with unsupervised feature design methods SDALF[18] and eBicov[34] to show the efficacy of our method. Shinpuhkan2014dataset has

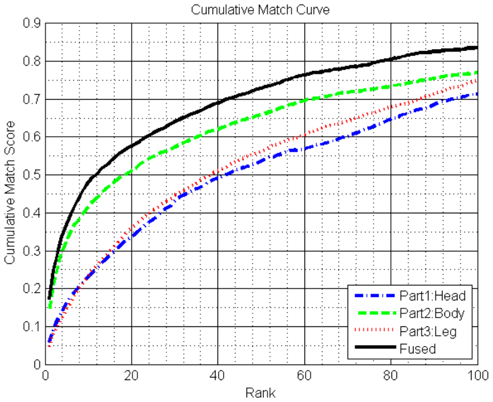

Fig. 8. The CMC curves on the VIPeR dataset, the training set is CASPR.

Table 1. Comparison with the DTRSVM and DML on VIPeR

Methods	Training sets	Rank1(%)	Rank10(%)	Rank20(%)	Rank30(%)
DTRSVM[28]	i-LIDS	8.26	31.39	44.83	53.88
DTRSVM[28]	PRID	10.90	28.20	37.69	44.87
DML[29]	CUHK Campus	16.17	45.82	57.56	64.24
Ours	CUHK02	16.90	45.89	58.58	65.32
Ours	CASPR	17.22	48.01	57.56	64.15
Oures	Shinpuhkan2014dataset	**18.64**	**48.54**	**60.63**	**68.48**

been chosen as our training set, we conduct experiments on VIPeR with the same data partition provided by SDALF, and also conduct experiments on i-LIDS with the same protocol of SDALF. The performances are summarized in Table 2. The results show that the person representation learned by our method has good generalization performance. Moreover, the proposed method directly extract feature from the original image without silhouette mask.

Table 2. Comparison with the unsupervised feature design methods

Methods	Test sets	Rank1(%)	Rank5(%)	Rank10(%)	Rank20(%)
SDALF[18]	VIPeR	19.87	38.89	49.37	65.73
eBiCov[34]	VIPeR	20.66	42.00	**56.18**	68.00
Ours	VIPeR	**21.27**	**43.99**	56.11	**69.72**
SDALF[18]	i-LIDS	28.49	48.21	57.28	68.26
Ours	i-LIDS	**29.83**	**50.76**	**61.04**	**73.64**

4.4 Results on PRID and i-LIDS

Since using Shinpuhkan2014dataset as training set can get the best result, we conduct two more cross dataset experiments on PRID and i-LIDS by the same models when testing the VIPeR dataset. PRID dataset consists of person images from two static surveillance cameras. Total 385 persons are captured by camera A, while 749 persons captured by camera B. The first 200 persons appeared in both cameras, and the remainders only appear in one camera. 100 out of the 200 image pairs are randomly taken out while and the others for testing. We repeat this for 10 times. The recognition rates are summarized in Table 3 as well as the comparison with DTRSVM.

Table 3. Comparison with the DTRSVM on PRID 2011

Methods	Training sets	Rank1(%)	Rank10(%)	Rank20(%)	Rank30(%)
DTRSVM[28]	VIPeR	4.6	17.25	22.9	28.1
DTRSVM[28]	i-LIDS	3.95	18.85	26.6	33.2
Ours	Shinpuhkan2014dataset	**16.80**	**43.30**	**52.40**	**56.80**

From the results we can see that the PRID is a very difficult dataset, the recognition rates of DTRSVM are very low. The chromatic aberration of the images in camera B folder might be the reason. However, our method still can get impressive results which outperform DTRSVM significantly.

The i-LIDS contains 476 person images from 119 persons, 80 persons are randomly chosen for testing. We choose one image from each person randomly to consist the gallery set, the remaining images are used as the probe set. This procedure is repeated 10 times. The CMC curves are shown in Figure 9, the recognition rates are summarized in Table 4 as well as the comparison with other state-of-the-art methods.

In Figure 9, we can see that the performance of the head part almost equals that of the fusion and the performance of the leg part is poor. The reason is that the i-LIDS dataset has many occlusions, the leg parts are often obscured by suitcases. The segmentation is not precise as other datasets, only the head parts are relatively stable. The results presented in Table 4 are exciting, our method outperforms most learning based methods without any training procedure of the testing set. Given all the cross dataset experiments we have conducted, we can see that the person representation learned by our method can perform well on different datasets without prior information of the testing sets.

5 Conclusions

The cross dataset person re-identification problem has not been paid much attention before, but it is very important for practical applications. This paper

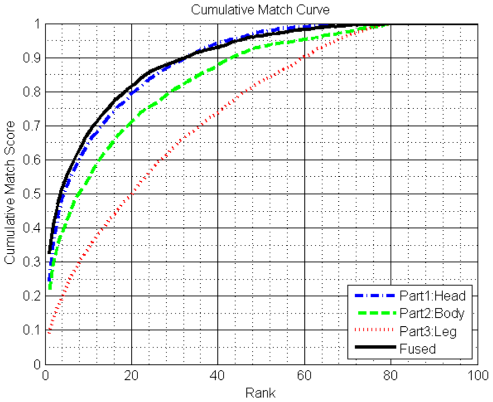

Fig. 9. The CMC curves on i-LIDS, the training set is Shinpuhkan2014dataset.

Table 4. Comparison with the state-of-the-art on i-LIDS with p = 80

Methods	Rank1(%)	Rank5(%)	Rank10(%)	Rank20(%)	Rank30(%)
MCC[35]	12.00	33.66	47.96	67.00	NA
ITM[8]	21.67	41.80	55.12	71.31	NA
Adaboost[24]	22.79	44.41	57.16	70.55	NA
LMNN[6]	23.70	45.42	57.32	70.92	NA
Xing's[36]	23.18	45.24	56.90	70.46	NA
L1-norm	26.73	49.04	60.32	72.07	NA
Bhat.	24.76	45.35	56.12	69.31	NA
PRDC[9]	**32.60**	54.55	65.89	78.30	NA
Ours	32.41	**55.19**	**67.70**	**81.54**	**88.81**

proposed a feature learning method by using CNN for the cross dataset person re-identification. The structure and training process were described in detail. Three public datasets, VIPeR, PRID and i-LIDS, were tested to evaluate the performance of the learned person representation. Extensive results illustrated that the learned representation had good generalization.

Acknowledgment

This work was supported by the Chinese National Natural Science Foundation Projects #61105023, #61103156, #61105037, #61203267, #61375037, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

References

1. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. (Conference on Computer Vision and Pattern Recognition (CVPR))
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1106–1114
3. Kawanishi, Y., Wu, Y., Mukunoki, M., Minoh, M.: Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In: Proc. of The 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV). (2014)
4. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: ACCV (1). (2012) 31–44
5. Li, W., Wang, X.: Locally aligned feature transforms across views. In: CVPR. (2013) 3594–3601
6. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. Advances in neural information processing systems **18** (2006) 1473
7. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian recognition with a learned metric. In: Computer Vision–ACCV 2010. Springer (2011) 501–512
8. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the 24th international conference on Machine learning, ACM (2007) 209–216
9. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR. (2011) 649–656
10. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2288–2295
11. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 3610–3617
12. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 498–505
13. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person Re-Identification. Springer (2014)
14. Hu, Y., Liao, S., Lei, Z., Yi, D., Li, S.Z.: (Exploring structural information and fusing multiple features for person re-identification)
15. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR (2). (2006) 1528–1535
16. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: ICDSC. (2008) 1–6
17. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.H.: Shape and appearance context modeling. In: ICCV. (2007) 1–8
18. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR. (2010) 2360–2367
19. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: Computer Vision–ECCV 2012. Workshops and Demonstrations, Springer (2012) 413–422

20. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC. Volume 2. (2011) 6
21. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 3586–3593
22. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching, ICCV (2013)
23. Liu, Y., Shao, Y., Sun, F.: Person re-identification based on visual saliency. In: Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on, IEEE (2012) 884–889
24. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV (1). (2008) 262–275
25. Prosser, B., Zheng, W.S., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: BMVC. Volume 1. (2010) 5
26. Li, W., Wang, X.: Locally aligned feature transforms across views. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 3594–3601
27. Liu, C., Loy, C.C., Gong, S., Wang, G.: Pop: Person re-identification post-rank optimisation. In: International Conference on Computer Vision. (2013)
28. Ma, A., Yuen, P., Li, J.: Domain transfer support vector ranking for person re-identification without target camera label information. In: Computer Vision (ICCV), 2013 IEEE International Conference on. (2013) 3567–3574
29. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: International Conference on Pattern Recognition. (2014)
30. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE. (1998) 2278–2324
31. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML. (2010) 807–814
32. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International workshop on performance evaluation of tracking and surveillance, Citeseer (2007)
33. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Proceedings of the 17th Scandinavian Conference on Image Analysis. SCIA'11, Berlin, Heidelberg, Springer-Verlag (2011) 91–102
34. Ma, B., Su, Y., Jurie, F., et al.: Bicov: a novel image representation for person re-identification and face verification. In: British Machive Vision Conference. (2012)
35. Globerson, A., Roweis, S.T.: Metric learning by collapsing classes. In: NIPS. (2005)
36. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.J.: Distance metric learning with application to clustering with side-information. In: NIPS. (2002) 505–512