Advances in Computer Science Research (ACRS), volume 54
International Conference on Computer Networks and Communication Technology (CNCT2016)

Multi-Scale Convolutional Network for Person Re-identification

Qiong WU*

Electronics and Communication Engineering, Zhengzhou University, China

Keywords: Deep learning, Person re-identification.

Abstract. In the last several years, methods with learning procedure held the state-of-the-art results for person re-identification (re-id) problem, especially the metric learning algorithm. Recently, with the success of deep learning methods on many computer vision tasks, researchers started to put their focuses on learning high-performance features. In this paper, we propose a method by fusing features learned from a multi-scale convolutional neural network and the traditional hand-crafted features, which improves the performance significantly. The Shinpuhkan2014dataset has been chosen as the training set, and we evaluate the performances of the proposed method on VIPeR, PRID and i-LIDS. Experiments show that our method outperforms the existing methods and even approaches the performances of the methods which have a training step on the testing sets.

Introduction

Person re-identification(re-id) is a challenging problem and has attracted more and more attention in recent years. It aims to associate identities of individuals across disjoint views in non-overlapping camera networks. 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, person re-identification problem is a difficult task. In practical applications, the person images are captured from surveillance cameras which 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, variations in illumination, view-point, background, pose, camera parameter and occlusion make the same person look differently under different camera views 1. For the existing re-id datasets, the existing problem is as follows:(1) lack of samples to represent the appearance variations of each individual; (2) the distribution of the intra classes and inter classes are unstable due to the diversities and ambiguities of the samples.

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Figure 1. (a) Example pairs of images from the VIPER database. (b) Example pairs of images from the i-LIDS database. Images in the same column represent the same person.

Like other computer vision tasks, the common solution to the re-id problem can be divided into two steps: feature extraction and classifier design. For the feature extraction step, the main goal is to extract effective features. However, it is hard to find visual features which are both distinctive and stable under various conditions. Another step aims to learn optimal distance measure for all features jointly via distance learning theory. It is an effective way when 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 this kind of methods have bad generalization and easy to over-fitting.

In this paper, we proposed a person re-identification method based on convolutional neural network (CNN). Compared with the common two-step re-id methods, CNN has learnable parameters and there is no gap between the feature extraction step and the classification step. In CNN, all steps are optimized together to minimize the estimation error. Furthermore, we also apply the traditional features to ensure the basic performance. The reason is that deep learning framework is especially suitable for dealing with large training sets, although the scales of re-id datasets are getting larger, they are still not comparable to the scales in other fields [25, 15], both in the number of subjects and the number of images per subject.

In order to dig out the power of CNN, we apply several strategies into the proposed method. 1) Divide the person image into many local patches by a pre-defined rule and feed them into CNN, which can make the proposed method more robust to image translation and pose variations. 2) Person image is cropped into many multi-scale patches and a regression function is learned on these patches jointly. Due to the complementary information between different parts and scales, multi-scale analysis can improve the performance of CNN significantly. 3) Extract traditional hand-crafted features on the cropped patches which do not be affected by the training data. Apply this kind of features as a supplement component can improve the generalization. Finally, we get a model by training a multi-scale CNN and fusing with the hand-crafted features to illustrate the flexibility of the proposed method.

Related Work

The recognition rate of person re-identification has increased a lot over the last several years. Among the state-of-the-art methods, metric learning (ML) approaches have played very important roles [27, 3, 2, 32, 14, 18, 10, 7]. Weinberger et al. [27] proposed the LMNN method to learn a Mahanalobis distance metric for k-nearest neighbor (kNN) classification by semidefinite programming. Davis et al. [2] 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. [32] 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. [14] proposed the KISSME method to learn a distance metric from equivalence constraints based on a statistical inference perspective. Li at al. [18] 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, which are not suitable for the practical applications.

Another type of methods tries to tackle the re-id problem by seeking feature representations which are both distinctive and stable under various circumstances [13, 5, 11, 26]. Farenzena et al. [4] 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 variations. Malocal et al. [21] turned the local descriptors into Fisher Vectors to produce a global representation of the image. Cheng et al. [1] utilized Pictorial Structures for person re-identification. Color information and color displacement within the whole body were extracted per part. Salience was also applied in person re-identification [31, 30, 19]. However, most of hand-crafted features are not distinctive and stable enough and may lose efficacy due to the new coming data.

Resent person re-identification methods usually take into account various aspects. Gray and Tao [9] applied AdaBoost to select good features out of a set of color and texture features. Prosser et al. [23] formulated the person re-identification problem as a ranking problem and applied the Ensemble RankSVM to learn a subspace where the potential true match gets the highest rank. In [17], visual features of an image pair from different views were first locally aligned by being projected to a common feature space and then matched with softly assigned metrics which are locally optimized.

In order to improve the performance in practical applications, some researchers started to conduct experiments on cross dataset person re-identification. Ma et al. [20] 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. Yi at al. [29] proposed a method called Deep Metric Learning (DML) which learned 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. 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.

Considering the obvious superiorities on large amount of training data and the development of computation resources, more and more researchers are carrying on studies on methods based on Deep Learning. Convolutional neural network has achieved great success in many computer vision tasks. For instance, DeepFace [24] and [16] have exhibited impressive results on face recognition and image classification. Therefore, we will address the person re-identification problem by constructing a multi-scale convolutional neural network as well as fusing the traditional hand-crafted features. In the following sections, the implementation details

of the proposed method will be described, and the comparison experiment results will be reported as well.

Proposed Method

The structure of the proposed method is shown in Fig.2, which includes many sub-networks for each patch. The details of the network are described in the following contents.

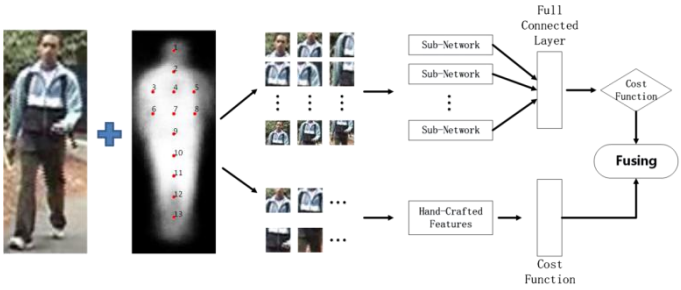

Figure 2. The flowchart of the proposed method. The input person image is cropped into many patches according to the average mask and landmarks. All patches are fed to the multi-scale convolutional network. Besides, hand-crafted features are extracted on the patches which located on the central axis of the body.

Body Patches

Cropping body patches are important for good person re-identification methods. Given the performances of most existing pedestrian alignment methods are not that good and very consuming, we propose a simple and effective way to get local patches. Firstly, we calculate the average pedestrian image of all the images from a dataset. In this paper, we choose the VIPeR dataset to get a mask and then resize it to 256×96. At last, we set 13 landmarks on the mask. The specific approach is to carry out uniform sampling on the central axis and more landmarks on the torso part, the sampling interval is 48 and 24 pixels on the torso part. The positions of the selected landmarks are shown in Fig.3.

Before cropping image patches, the resolution of mask and image are normalized to 256×96 and 128×48. All landmarks are transformed along with the normalization of images. On the normalized images, several 48×48 patches are cropped in 2 scales by taking the landmarks as center. The numbers of landmarks used in 2 scales are 13 and 8 from to small to large scale, therefore giving 21 multi-scale patches and the numbers can be seen in Fig.3. In large scale, we utilize more patches to capture the local information, meanwhile in small scale, the patches we choose will have a lot of overlap.

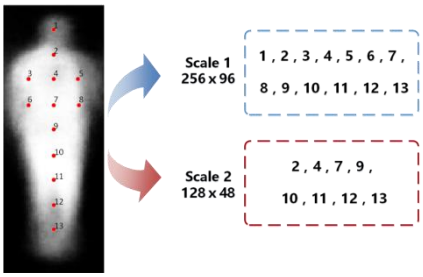

Figure 3. For each pedestrian image, 21 multi-scale patches are cropped according to the landmarks on the average mask. The resolution of all patches is 48*48.

The Architecture of CNN

As the architecture of the proposed network shown in Fig.2, the details of each layer are described in Fig.4. For the 21 groups of image patches, we create 21 sub-networks to process them respectively and fuse their responses in the final full connected layer to identify the person. This structure has two benefits: 1) 21 sub-networks can learn the particular features for each patch; 2) The final layer connects all sub-networks together, which can make them mutually complementary. Note that the parameters of the 21 sub-networks and the final layer in Fig.2 and Fig.4 are optimized in a whole process.

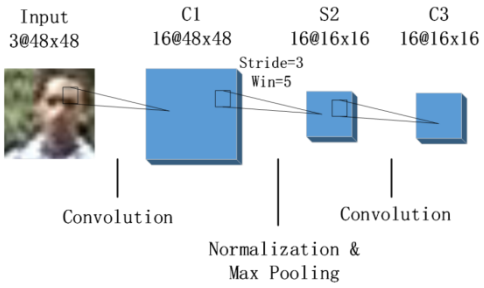

Figure 4. The structure of sub-network for each patch. The input of sub-network is pedestrian patches and the output is sent to the fully connected layer in Fig.2.

The sub-network for each patch is composed by a convolutional layer, a max pooling layer and a local layer (Locally-connected layer with unshared weights). The numbers of channels of the convolutional, pooling and local layers are both 16. Before convolution the input is padded by zero values, therefore the output has the same size with input. The filter size of C1 layer is 7×7 and the filter size of L3 layer is 3×3. ReLU neuron [15] is used as activation function for C1 and L3. The stride of S2 is 3 that means it down-samples the feature maps from 48×48 to 16×16 and S2 includes a cross-channel normalization unit. The output of the sub-network (L3) has 16×16×16 = 4096 dimensions. Therefore the input of F4 has 4096×21 = 86016 dimensions. F4 uses square difference as cost function, therefore it can be seen as a linear regression layer. In practice, we should pay attention to the magnitude of F4's output. Generally, we need introduce a scale factor to make them in the same order of magnitude. The network are optimized by Stochastic Gradient Descent (SGD).

Hand-Crafted Features

Due to the small scale of training set in re-id problem, the model trained by the CNN may not have good generalization ability, especially when the sample distribution of the testing set is different from the training set. Therefore, we extract hand-crafted features to assist. In this paper, we extract weighted color histograms (in HSV color space) by taking into account the pixel distance to the central axis.

As weighted color histograms extracted on image patches can capture the local information well, we apply the same cropping way introduced in Fig.3. The difference is that we extract hand-crafted features on the 8 patches in scale 2. In order to make the pixels near the central axis more important, we apply a weight for each pixel when extract the color histograms to avoid the pose variation as far as possible.

Then, weighted color histograms are built by applying a Gaussian kernel within each patch:

$$\mathrm{H(i)} = \sum_{x,y} \omega(x,y) s(I(x,y) \in B(i)) \tag{1}$$

where H represents the histogram, $s(.)$ is a bool function, $B(i)$ is the value range of the ith bin, $\omega(x,y)$ is calculated as follows:

$$\omega(x,y) = \frac{1}{Z}\exp[-(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2})] \quad (2)$$

where (x_0, y_0) is the center of the patch, σ_x and σ_y is the deviation parameter (we set σ_x and σ_y three times of the width and height of the patch, respectively).

Experiment

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

Four datasets are involved in the experiment. Shinpuhkan2014dataset is chosen as the training sets. Then, we evaluate the performances on VIPeR[8], PRID[12] and i-LIDS[32].

Training

Shinpuhkan2014dataset consists of more than 22,000 images of 24 people which are captured by 16 cameras installed in a shopping mall "Shinpuhkan". All images are manually cropped and resized to 128×48 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 Fig.5.

Figure 5. 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.

Results on VIPeR

VIPeR 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.

For the VIPeR dataset, we split it 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. 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 evaluation accuracies of person re-identification are currently very low. Our method gets very impressive results and even approaches the performance of some methods in

single database setting, such as ELF [9] and PRDC [32]. Besides, we can see that 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.

Table 1. Comparison with the DTRSVM and DML on VIPeR

Methods	Training sets	Rank1(%)	Rank10(%)	Rank20(%)	Rank30(%)
DTRSVM[20]	i-LIDS	8.26	31.39	44.83	53.88
DTRSVM[20]	PRID	10.90	28.20	37.69	44.87
DML[29]	CUHK Campus	16.17	45.82	57.56	64.24
Ours	Shinpuhkan2014dataset	**18.96**	**48.84**	**61.63**	**69.42**

We also compared the proposed method with unsupervised feature design methods SDALF[4] and eBicov[22] to show the performance of the extracted feature. 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[4]	VIPeR	19.87	38.89	49.37	65.73
eBiCov[22]	VIPeR	20.66	42.00	56.18	68.00
Ours	VIPeR	**21.73**	**44.26**	**56.71**	**70.34**

Results on PRID and i-LIDS

PRID 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.

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.

Since using Shinpuhkan2014dataset is suitable to be a training set, we conduct two more cross dataset experiments on PRID and i-LIDS by the same model when testing the VIPeR dataset. PRID dataset consists of person images from two static surveillance cameras. We repeat the test for 10 times and calculate the average performance. 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[20]	VIPeR	4.6	17.25	22.9	28.1
DTRSVM[20]	i-LIDS	3.95	18.85	26.6	33.2
Ours	Shinpuhkan2014dataset	**16.92**	**44.23**	**53.64**	**57.24**

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.

For the i-LIDS dataset, we also repeat the testing procedure 10 times. The recognition rates are summarized in Table 4 as well as the comparison with other state-of-the-art methods.

Table 4. Comparison with methods on i-LIDS with p = 80

Methods	Rank1(%)	Rank5(%)	Rank10(%)	Rank20(%)
MCC[6]	12.00	33.66	47.96	67.00
ITM[2]	21.67	41.80	55.12	71.31
Adaboost[9]	22.79	44.41	57.16	70.55
LMNN[27]	23.70	45.42	57.32	70.92
Xing's[28]	23.18	45.24	56.90	70.46
L1-norm	26.73	49.04	60.32	72.07
Bhat.	24.76	45.35	56.12	69.31
PRDC[32]	32.60	54.55	65.89	78.30
Ours	**32.74**	**55.43**	**67.97**	**82.44**

We can see 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.

Conclusion

This paper proposes a method by fusing features learned from a multi-scale CNN and the traditional hand-crafted features. The Shinpuhkan2014dataset has been chosen as the training set, and we evaluate the performances of the proposed method on VIPeR, PRID and i-LIDS. Experiments show that our method outperforms the existing methods.

References

[1] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In BMVC, volume 2, page 6, 2011. 2

[2] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In Proceedings of the 24th international conference on Machine learning, pages 209–216. ACM, 2007. 2, 6

[3] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In Computer Vision–ACCV 2010, pages 501–512. Springer, 2011. 2

[4] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In CVPR, pages 2360–2367, 2010. 2, 4, 5

[5] N. Gheissari, T. B. Sebastian, and R. Hartley. Person re-identification using spatiotemporal appearance. In CVPR (2), pages 1528–1535, 2006. 2

[6] A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In NIPS, 2005. 6

[7] S. Gong, M. Cristani, S. Yan, and C. C. Loy. Person Re-Identification. Springer, 2014. 2

[8] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In IEEE International workshop on performance evaluation of tracking and surveillance. Citeseer, 2007. 4

[9] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In ECCV (1), pages 262–275, 2008. 2, 4, 6

[10] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In Computer Vision, 2009 IEEE 12th International Conference on, pages 498–505. IEEE, 2009. 2

[11] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In ICDSC, pages 1–6, 2008. 2

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In Proceedings of the 17th Scandinavian Conference on Image Analysis, SCIA'11, pages 91–102, Berlin, Heidelberg, 2011. SpringerVerlag. 4

[13] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Exploring structural information and fusing multiple features for person re-identification. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, pages 794–799. IEEE, 2013. 2

[14] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2288–2295. IEEE, 2012. 2

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012. 1, 3

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1106–1114, 2012. 2

[17] W. Li and X. Wang. Locally aligned feature transforms across views. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 3594–3601. IEEE, 2013. 2

[18] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 3610–3617. IEEE, 2013. 2

[19] Y. Liu, Y. Shao, and F. Sun. Person re-identification based on visual saliency. In Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on, pages 884–889. IEEE, 2012. 2

[20] A. Ma, P. Yuen, and J. Li. Domain transfer support vector ranking for person re-identification without target camera label information. In Computer Vision (IC-CV), 2013 IEEE International Conference on, pages 3567–3574, Dec 2013. 2, 5, 6

[21] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In Computer Vision–ECCV 2012.Workshops and Demonstrations, pages 413–422. Springer, 2012. 2

[22] B. Ma, Y. Su, F. Jurie, et al. Bicov: a novel image representation for person re-identification and face verification. In British Machive Vision Conference, 2012.4, 5

[23] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In BMVC, volume 1, page 5, 2010. 2

[24] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Conference on Computer Vision and Pattern Recognition (CVPR). 2

[25] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1701–1708, 2014. 1

[26] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In ICCV, pages 1–8, 2007. 2

[27] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. Advances in neural information processing systems, 18:1473, 2006. 2, 6

[28] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell. Distance metric learning with application to clustering with side-information. In NIPS, pages 505–512, 2002. 6

[29] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In International Conference on Pattern Recognition, 2014. 2, 5

[30] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. ICCV, 2013. 2

[31] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 3586–3593. IEEE, 2013. 2

[32] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In CVPR, pages 649–656, 2011. 2, 4, 6