3A: A PERSON RE-IDENTIFICATION SYSTEM VIA ATTRIBUTE AUGMENTATION AND AGGREGATION

Zheng Wang[1], *Ruimin Hu*[1,*], *Chao Liang*[1], *Rui Shao*[1], *and Yi Yu*[2]

[1]State Key Laboratory of Software Engineering, School of Computer, Wuhan University, China
[2]Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan

ABSTRACT

Person re-identification is a key technique to match person images captured in non-overlapping camera views. Due to the sensitivity of low-level visual features to viewpoint change, scale zooming and illumination variation, high-level semantic attributes, more stable to the environmental change, begins to be investigated to improve the robustness of the representation. However, confusions may occur caused by the limited expression ability of coarse-grained semantic attributes. To explore the attribute value, we introduce a new framework including two steps: (1) attribute augmentation, by mining non-semantic attributes to supplement semantic attributes, and (2) attribute aggregation, by merging attribute-based ranking results into traditional feature-based ranking results. Experiments conducted on public datasets have validated the effectiveness of the proposed framework.

Index Terms— person re-identification, non-semantic attributes, attribute augmentation, attribute aggregation

1. INTRODUCTION

Person re-identification (re-id) is a task of matching persons across non-overlapping camera views [1]. Treated as a special image retrieval problem, re-id is attracting increasing attentions in the field of signal processing [2–4]. Relevant works can be generally categorized into two classes: metric learning and feature representation. The former emphasizes on seeking a proper measure to reflect the identity consistency among person images [5–7]. The latter aims at constructing expressive and robust feature descriptions [8–12]. Most of feature representation approaches have relied on low-level visual features [8], which are expressive but sensitive to viewpoint change, scale zooming and illumination variation [13–17]. In comparison, sematic attribute based vectors are relatively robust to person's appearance change [18]. To this end, some researchers start to investigate how to improve image representation with semantic attributes: Ryan Layne *et al.* [18] proposed to learn a selection and weighting of mid-level semantic attributes to describe people appearance. Cheng-Hao Kuo *et al.* [19] applied semantic color names to describe a person image. Yubin Deng *et al.* [20] released a new pedestrian attribute dataset, which is by far the largest and most diverse of its kind.

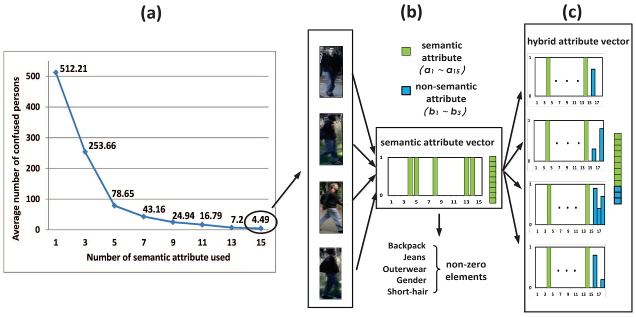


Fig. 1. (a) Tendency of average number of confused persons in classification with increase semantic attributes used, which are from [18] denoted as $(a_1, .., a_{15})$, such as ”shorts”, ”backpack”, ”long-hair” and so on. (b) An example of 4 confused persons in classification. (c) A hybrid representation augmented by a non-semantic part $(b_1, .., b_3)$.

However, two characteristics for person re-identification task were ignored by these semantic attribute based approaches: (1) ***Coarse-grained.*** Such semantic attributes used for person re-identification is pre-defined, whose quantity is far from sufficient in forming a expressive space. Take Fig.1(b) as an example, 15 semantic attributes, proposed by [18], are used to represent person appearance. Even with all the ground truth semantic attributes, some persons still share the same attribute vectors, leading to a theoretically indistinguishable result. (2) ***Samples-deficiency.*** As shown in Fig.1(a), with the increase number of attributes used, the average number of confused persons decreases. Therefore, enriching semantic attributes is an effective way to improve person re-id performance. However, training new attribute detector needs massive annotated samples, which are usually unavailable in the practical application. Consequently, we need to find a method

[*] R. Hu is the corresponding author, and is also with the National Engineering Research Center for Multimedia Software, the Collaborative Innovation Center of Geospatial Technology, and the Hubei Provincial Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University.

which augments the original attribute vectors without any additional annotated samples.

Inspired by [21], we propose to augment semantic attribute vectors with non-semantic attributes, forming a hybrid attribute representation. Compared to the robust but inexpressive semantic attributes, visual features are more expressive. Hence, we attempt to mine useful information from visual features, and generate non-semantic attributes. An *autoencoder* [22] model is adopted, which learns a latent space by minimizing the loss in the reconstruction of input space. For our method, this model consists of two steps: *encoding* and *decoding*. After attribute augmentation, as shown in Fig.1(c) non-semantic attributes have effectiveness to supplement semantic attribute vectors with more expression power. This augmented part is obtained by efficiently mining under-exploited information from existing visual features, without any more training samples.

Attribute vectors are robust refined with more expressive dimensions. However, there may still exist useful expressive information for the low-level feature vectors, especially for the feature descriptors that have not been exploited in the attribute augmentation step. Based on this consideration, attribute aggregation is proposed to mine more expressive information in traditional feature-based ranking list to further improve the performance of hybrid attribute based method.

We name these two steps as **3A** framework. The main contribution of this paper can be summarized as follows: (1) A hybrid attribute representation is proposed by attribute augmentation, in which the representation performance of attribute vectors can be developed with non-semantic part. (2) A refined ranking list is obtained by attribute aggregation, in which the attribute-based ranking list is aggregated with traditional feature-based ranking list. Experiments conducted on VIPeR [23] and PRID [24] datasets have shown the effectiveness of the proposed method.

2. OUR APPROACH

2.1. Overview

An overview of our approach is shown in Fig.2. It can be divided into two main parts: attribute augmentation and attribute aggregation. The former aims at augmenting the original semantic attribute representation with non-semantic part. The model *autoencoder* learns latent space by minimizing the loss in the reconstruction of input space, which is used to generate non-semantic part in latent space, forming a hybrid attribute vector. After attribute augmentation, an attribute-based ranking list is obtained. The latter aims at refining it by aggregating expressive information in traditional feature-based ranking list. The similarity ranking aggregation (SRA) method fully exploits the complementarity information of different baseline methods for ranking aggregation, and generates a refined ranking list.

2.2. Attribute Augmentation

Given K person images, and the i-th of them has a feature vector $x_i \in \mathbb{R}^d$ and a semantic attribute vector $a_i \in \mathbb{R}^n$. All people form a feature data matrix $X \in \mathbb{R}^{d\times K}$. Our goal is to augment a non-semantic part $b_i \in \mathbb{R}^m$, forming a hybrid attribute vector $[a_i,\ b_i]$ for each person, where [.] denotes the concatenation of vectors. To this end, the model *autoencoder* [22] is adopted, which aims at augmenting semantic attribute representation with non-semantic part. This model has the following two characteristics: (1) information in the input feature vector is preserved in the reconstructed future vector as much as possible; (2) the non-semantic part in hybrid attribute vector is learned automatically instead of learned by classifiers. This is achieved by a two-layered construction: encoding layer and decoding layer. In the first layer, this model encodes the input feature vector into hybrid attribute vector, which is composed of two parts: a known semantic part obtained by existing methods and a non-semantic part learned from visual features by an encoding function. In the second layer, we try to reconstruct feature vector from the joint two parts in the hybrid attribute vector, which means that the non-semantic part only needs to encode the information that the semantic part lacks. Consequently, we have found a simple way to reveal the lacking expressive information in semantic attribute vector as non-semantic attribute.

Encoding process. Firstly, each component of semantic attribute vector a_i is obtained from prediction of trained SVM attribute classifiers. An training image dataset is exploited to train attribute classifiers, which consists of visual feature vectors extracted on samples, along with the pre-labeled attributes contributed by [20]. Attribute will be labeled +1, if it appears in the image, and -1 otherwise. For each kind of attribute, a linear SVM classifier is trained separately with the software package Vlfeat [25]. Secondly, an encoding function E encodes feature vector x_i only for the non-semantic attribute part as $b_i = S(W_B x_i)$, where $W_B \in \mathbb{R}^{m\times d}$ is the augmentation matrix containing all the attribute augmentation parameters, $S(z) = 1/(1 + exp(-z))$ is a sigmoid function which ensures values of b_i are in a range comparable to the a_i. This process is shown as

$$[a_i, b_i] = [a_i, E(x_i)] = [a_i, S(W_B x_i)] \tag{1}$$

Decoding process. The decoding function D aims at reconstructing the image to its original input feature space x_i from hybrid attribute vector$[a_i,\ b_i]$. This process is shown as Eq.2, where U is the reconstruction matrix decomposed as $U_A \in \mathbb{R}^{d\times n}$ and $U_B \in \mathbb{R}^{d\times m}$.

$$\hat{x}_i = D([a_i, b_i]) = U[a_i, b_i] = U_A a_i + U_B b_i \tag{2}$$

Reconstruction loss. The reconstruction loss measures the loss incurred in the reconstruction of input feature vectors of all people, which can be used to guide the learning of W_B and U. All people form a semantic attribute matrix $A \in$

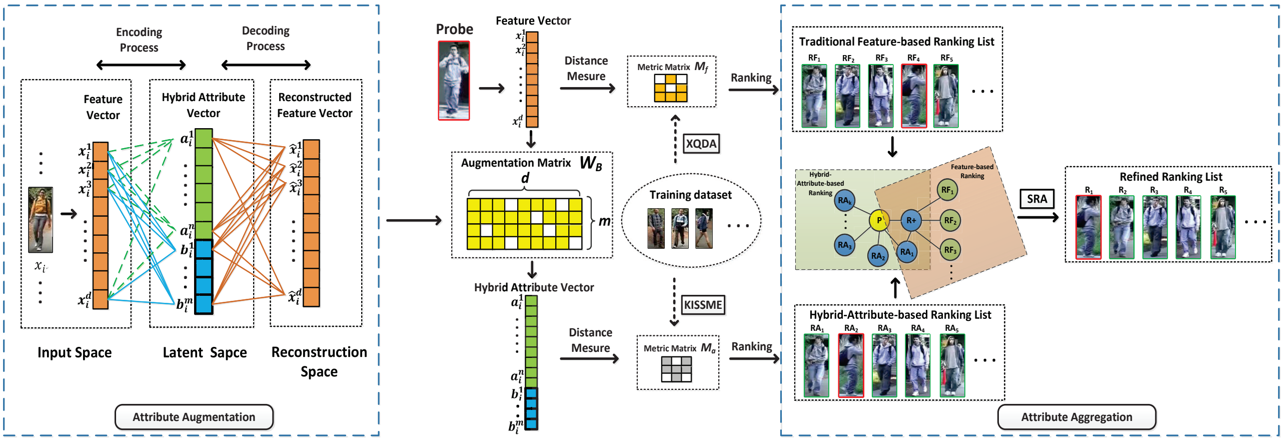


Fig. 2. An overview of the proposed method. In step of Attribute Augmentation, an augmentation matrix $\boldsymbol{W_B}$ is generated which can be used to produce a hybrid attribute vector. In step of Attribute Aggregation, a refined ranking list is obtained.

$\mathbb{R}^{n\times K}$ and a non-semantic attribute matrix $B\in \mathbb{R}^{m\times K}$. We use a squared error loss [26] as Eq.3, where $\hat{X} = U_A A+U_B B$ denotes the whole reconstruction representations of X, and $\|.\|$ denotes Frobenius norm of a matrix.

$$L_R = \sum_{i=1}^{K} \|x_i - \hat{x}_i\|^2 = \|X - \hat{X}\|^2 \tag{3}$$

Parameters Optimization. Minimizing Eq.3 is a non-convex optimization problem. An alternating optimization is adopted in our solution, which is optimizing one matrix at a time while fixing the others. Concretely, when the matrices W_B, U_B in Eq.3 are fixed, we obtain the closed form solution for updating matrix U_A by solving a ridge regression problem:

$$\min_{U_A} \parallel U_A A + U_B B - X \parallel^2 + \alpha \parallel U_A \parallel^2 \tag{4}$$

where α is the trade-off variables. And we did the same as U_B. After updating U_A, U_B, Broyden-Fletcher-Goldfarb-Shanno gradient descent method with limited-memory variation (L-BFGS) is used for updating the matrix W_B. After the convergence of objective function, we can get the final results of W_B.

2.3. Attribute Aggregation

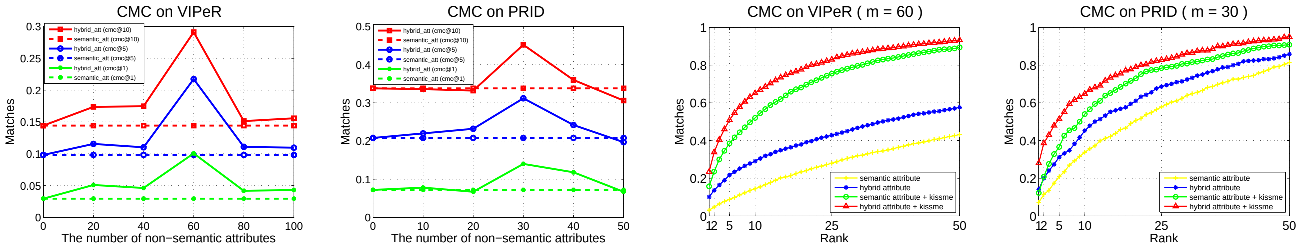
After augmented with more expressive information in attribute augmentation, attribute vectors are still much more low-dimensional compared to traditional visual feature vectors. Consequently, the feature-based ranking list contains much other useful expressive information attribute-based ranking list lacks. Base on these reasons, attribute aggregation is proposed to aggregate attribute-based ranking list with expressive information in traditional feature-based ranking list.

Concretely, for a strongly similar gallery to probe based on hybrid attributes (appearing in top-k ranking list), if it is the strongly similar one based on traditional visual features as well, we push it forward in original ranking list, which forms a refined ranking list. The method SRA [27] is adopted to realize attribute aggregation.

For example, a traditional feature-based ranking list RF is generated, where the superscript indicates the ranking orders. Meanwhile, after attribute augmentation, a attribute-based ranking list RA is obtained. As the attribute aggregation illustrated in Fig.2, the center is the probe while the nodes are the galleries in ranking list based on two methods. For this probe, strongly similar galleries are obtained from the intersection set of top-k results achieved by two methods, from which images in attribute-based ranking list are firstly selected, denoted as R_+. Then they are regarded as the new probes to requery by feature-based method, which leads to the generation of the refined ranking list.

3. EXPERIMENT

The experiment settings are described as follows. (1) ***Datasets.*** As most existing person re-id researches did, two publicly representative datasets, the VIPeR dataset [23] and the PRID [24] dataset are adopted to verify our approach. We choose these datasets as they provide many challenges faced in practical surveillance, *i.e.*, viewpoint, pose and illumination changes, different backgrounds, occlusions, etc. The VIPeR dataset contains 632 persons, each of which has two images normalized to 128×48 pixels. The set of 632 image pairs were randomly split into two sets of 316 image pairs each, one for training, while the other one for testing. For the PRID dataset, 400 shots of the first 200 person from each view of the single shot version are adopted to carry out the experiments. The images are scaled to 128×64 pixels. (2) ***Evaluation.*** Cumulative Matching Characteristic (CMC) curves [28] were used to calculate the average performance, and the value of CMC@k indicates the percentage of the real match ranked in the top k. The entire evaluation procedure was repeated 10 times. (3) ***Features.*** In this paper, the Local Maximal Occurrence (LOMO) features [29] were extracted to train the SVM semantic attribute classifiers. Principal



(a) (b)

Fig. 3. (a) Performances of hybrid attribute vectors augmented by different number of non-semantic attributes on VIPeR and PRID datasets. (b) The best attribute augmentation performance on VIPeR and PRID datasets.

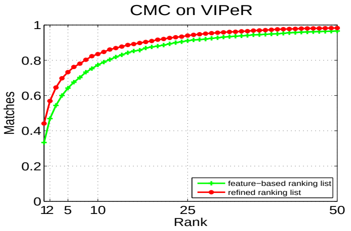


Fig. 4. Attribute aggregation performance on VIPeR dataset.

Table 1. Comparable results with the-state-of-the-art person re-id methods of CMC values (%) on VIPeR dataset.

Method	rank@1	10	25	50
W. AIR [18]	17.40	50.84	74.44	86.44
SCN [19]	23.9	56.2	73.1	86.5
SCNCD [31]	20.7	60.6	79.1	90.4
eSDC-ocsvm [32]	26.7	62.4	-	-
LADF [33]	13.5	56.01	79.64	92.51
KISSME [30]	19.6	62.2	80.7	91.8
Z.Shi, et al. [34]	31.1	82.8	**94.9**	-
Improved Deep [35]	34.81	75.63	84.49	-
Improved NFST [36]	42.28	82.94	92.06	-
Ours	**44.11**	**83.51**	93.99	**98.29**

component analysis was applied to reduce the feature dimensionality to 400 for VIPeR and 125 for PRID. (4) ***Attributes.*** The semantic attributes we adopted were derived from PEdesTrian Attribute (PETA) dataset [20]. This dataset contains 61 types of attributes. Since some of the attributes rarely appear, we selected 35 types of attributes for VIPeR dataset, such as ”personalFemale”, ”personLess30”, ”carryingbackpack”. For the PRID dataset, we selected 14 types of common semantic attributes. (5) ***Metrics.*** For better measuring the augmented attribute vector similarity as well, KISSME [30] is adopted, in which a Mahalanobis distance metric matrix M_a is learned for computing the distance between the attribute vectors. Meanwhile, the Quadratic Discriminant Analysis (XQDA) method [29] was utilized to learn a metric matrix M_f used for distance measure between traditional feature vectors.

The results are shown in Fig.3 and Fig.4. Several conclusions can be drawn as follows. (1) The number of non-semantic attributes is a key parameter affecting the effectiveness of attribute augmentation. With non-semantic attributes augmented increasing, we compared CMC@1,5,10 between semantic attribute vectors and hybrid attribute vectors respectively. Concretely, as shown in Fig.3(a) for the VIPeR, attribute augmentation shows only minute improvements over semantic attribute representation when augmenting with small dimensions, such as m = 20 or 40. We consider the reason is that the expression power of small additional dimensions is overwhelmed by the strong semantic prior. Whereas augmented with higher dimensions m = 60, hybrid attribute vectors becomes clearly better in comparison to the semantic attribute representation. When augmenting with too large dimensions, such as m = 80 or 100, too much redundant information in non-semantic attributes leads to sensitivity, which worsens the whole representation performance of hybrid attribute vectors. From Fig.3(b) shows a similar tendency for the PRID dataset. (2) Attribute aggregation is satisfying. As illustrated in our experiments shown in the Fig.4, the refined ranking list is better than the feature-based ranking list. Furthermore, Table.1 summarizes the comparison results with the state-of-the-art re-id methods on the widely used VIPeR dataset. This table shows that our method achieves the best performance compared with the state-of-the-art methods.

4. CONCLUSION

In this paper, we address person re-identification problem via attribute augmentation and aggregation. A hybrid attribute representation is proposed by attribute augmentation, and a refined ranking list is then proposed by attribute aggregation. Extensive experiments show the superiority of our proposed framework.

5. ACKNOWLEDGMENT

The research was supported by National High Technology Research and Development Program of China (2015AA016306), Natural Science Foundation of China (61231015, 61671336, 61671332, 61562048), Natural Science Foundation of JiangSu Province (BK20160386), Natural Science Foundation of Hubei Province (2016CFB573), Technology Research Program of Ministry of Public Security (2016JSYJA12), EU FP7 QUICK project (PIRSES-GA-2013-612652), and Jiangxi Youth Science Foundation of China (20151BAB217013).

6. REFERENCES

[1] Zheng Wang, Ruimin Hu, Yi Yu, Chao Liang, and Wenxin Huang, "Multi-level fusion for person re-identification with incomplete marks.," in *ACM MM*, 2015.

[2] Cong Ma, Zhenjiang Miao, and Min Li, "Saliency preprocessing for person re-identification images," in *ICASSP*, 2016.

[3] J. Oliver, Antonio Albiol, Antonio Albiol, and Jose Manuel Mossi, "Re-identifying people in the wild," in *ICASSP*, 2013.

[4] Yanna Zhao, Xu Zhao, and Yuncai Liu, "Person re-identification by free energy score space encoding," in *ICIP*, 2014.

[5] M Kostinger, Martin Hirzer, Paul Wohlhart, and et al, "Large scale metric learning from equivalence constraints.," in *CVPR*, 2012.

[6] Li Z, Chang S, Liang F, and et al, "Learning locally-adaptive decision functions for person verification.," in *CVPR*, 2013.

[7] Jin Wang, Zheng Wang, Changxin Gao, Nong Sang, and Rui Huang, "Deeplist: Learning deep features with adaptive listwise constraint for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, 2016.

[8] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010.

[9] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification.," in *CVPR*, 2013.

[10] Le An, Xiaojing Chen, Songfan Yang, and Xuelong Li, "Person re-identification by multi-hypergraph fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, 2016.

[11] Le An, Mehran Kafai, Songfan Yang, and Bir Bhanu, "Person reidentification with reference descriptor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 776–787, 2016.

[12] Le An, Xiaojing Chen, Songfan Yang, and Bir Bhanu, "Sparse representation matching for person re-identification," *Inform. Sciences*, vol. 355, pp. 74–89, 2016.

[13] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, "Person re-identification by probabilistic relative distance comparison.," in *CVPR*, 2011.

[14] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu, "Person re-identification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2553, 2016.

[15] Zheng Wang, Ruimin Hu, Chao Liang, and Yi Yu, "Zero-shot person re-identification via cross-view consistency," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.

[16] Junjun Jiang, Ruimin Hu, Zhongyuan Wang, and Zhen Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1268–1281, 2014.

[17] Junjun Jiang, Ruimin Hu, Zhongyuan Wang, and Zhen Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4220–31, 2014.

[18] Ryan Layne, Timothy Hospedales, and Shaogang Gong, "Person re-identification by attributes," in *BMVC*, 2012.

[19] Cheng-Hao Kuo, Sameh Khamis, and Vinay Shet, "Person re-identification using semantic color names and rankboost," in *WACV*, 2013.

[20] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *ACM MM*, 2014.

[21] Felix X. Yu, Rongrong Ji, Ming-Hen Tsai, Guangnan Ye, and Shih-Fu Chang, "Weak attributes for large-scale image retrieval.," in *CVPR*, 2012.

[22] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H. Lampert, "Augmented attribute representations," in *ECCV*, 2012.

[23] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking.," in *PETS*, 2007.

[24] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification.," in *Image Analysis*, 2011.

[25] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms.," in *ACM MM*, 2010.

[26] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[27] Mang Ye, Jun Chen, Qingming Leng, Chao Liang, Zheng Wang, and Kaimin Sun, "Coupled-view based ranking optimization for person re-identification.," in *MMM*, 2015.

[28] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling.," in *ICCV*, 2007.

[29] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li, "Person re-identification by local maximal occurrence representation and metric learning.," in *CVPR*, 2015.

[30] Martin Kostinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints.," in *CVPR*, 2012.

[31] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification.," in *ECCV*, 2014.

[32] Z. Rui, O. Wanli, and W. Xiaogang, "Unsupervised salience learning for person re-identification.," in *ICCV*, 2013.

[33] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith, "Learning locally-adaptive decision functions for person verification.," in *ICCV*, 2013.

[34] Zhiyuan Shi, Timothy M. Hospedales, and Tao Xiang, "Transferring a semantic representation for person re-identification and search.," in *CVPR*, 2015.

[35] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification.," in *CVPR*, 2015.

[36] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a discriminative null space for person re-identification.," in *CVPR*, 2016.