Specific Person Retrieval via Incomplete Text Description

Mang Ye[1,2], Chao Liang[1,2,*], Zheng Wang[1,2], Qingming Leng[3], Jun Chen[1,2], Jun Liu[1,2]

[1]National Engineering Research Center for Multimedia Software,
School of Computer, Wuhan University, Wuhan, 430072, China
{yemang,cliang,wangzwhu,chenj,jliu_newbee}@whu.edu.cn
[2]Research Institute of Wuhan University in Shenzhen, China
[3]School of Information Science & Technology, Jiujiang University, China
lengqingming@126.com

ABSTRACT

Searching for specific persons from surveillance videos captured by different cameras, is a key yet under-addressed challenge in multimedia system. Related person retrieval works mainly focus on searching person by visual appearance, known as person re-identification. However, the initial visual image may not be available in some practical applications. For example, the criminal is described by a text description indirectly, "*A young woman wearing a red casual with a backpack*", the traditional methods can not conquer this issue. Based on a set of pre-defined attributes that the text description query can be transformed to an attribute vector, thus can be used to retrieval in the gallery set. And yet, the user-provided attributes are sometimes incomplete. This new issue is defined as Specific Person Retrieval via Incomplete Text Description. In this paper, we conduct a specific attribute completion to enrich the original text query and generate a more expressive attribute vector. Then, a pairwise-based metric learning is introduced for completed attribute vectors. Extensive experiments conducted on two benchmark datasets have shown our superior performance.

Categories and Subject Descriptors

J.m [**Computer Applications**]: Miscellaneous;
H.3.3 [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL— *Retrieval models*

General Terms

Theory

Keywords

Person Retrieval, Attribute, Attribute Completion

1. INTRODUCTION

Specific person retrieval for video surveillance attracts a lot of attention in multimedia retrieval system [4]. Most of the existing approaches exploit person's visual appearance (single or multiple images of the query person) to search across different non-overlapping cameras, known as person re-identification [7,16,17]. However, in some practical surveillance application, merely text descriptions rather than visual appearances are available. For example, the witness told the police that the appearance of the criminal by oral text description, "*A young woman wearing a red casual with a backpack*", as shown in Fig. 1. Traditional person re-identification methods can't handle this issue, where the visual appearance of the query person is not obtainable. We named this novel issue as *specific person retrieval via text description*.

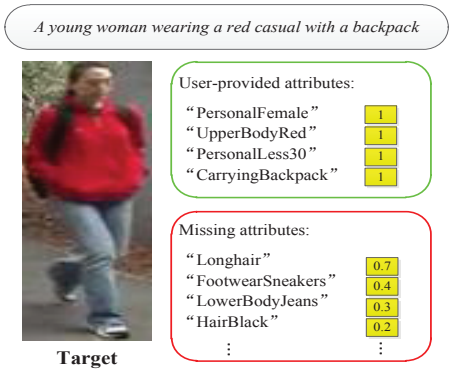


Figure 1: An illustration example of user-provided attributes. It can be seen that the user-provided attributes are incomplete. The attributes are derived from [3]. And the missing attributes are partly completed, where the value of each attribute denotes the owning probability..

In this condition, it is critical to solve this specific person retrieval problem merely by given specific attributes, which can be abstracted from the oral text description [10]. Related person re-identification have addressed some attribute-based methods. Layne *et al.* [8] learned a selection and weighting of mid-level semantic attributes to describe people. Liu *et al.* [10] propose a novel Attribute Restricted Latent Topic Model (ARLTM) to encode targets into semantic topics. Nguyen *et al.* [1] proposed an approach to exploit relationships between attributes for refining attribute detection results. Nevertheless, the above approaches are not proper for current issue for the reason that their attributes are originated from the person appearance, where the probe image is unavailable.

*indicates corresponding author

ICMR'15, June 23–26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749347.

Furthermore, the user-provided attributes are usually incomplete, as revealed in [15]. Specially, 61 attributes are predefined for person description in [3], while the user-provided text description can not cover all the pre-defined attributes, as illustrated in Fig.1. The incompleteness of user-provided attributes probably leads to performance degradations for person retrieval [10]. In a wider range level, attribute completion is widely conducted in traditional TBIR (tag-based image retrieval) [2,15]. Motivated by [9], we conduct a linear sparse reconstruction to complete the incomplete attributes. Specially, the original user-provided attribute vector is reconstructed with the training ones under constrains of sparsity. To better measuring the attribute vector similarity, a pairwise-based metric learning framework is firstly conducted for attributes.

In this paper, we present a novel method to solve the above new problem, *i.e.* person retrieval via incomplete text description in surveillance camera network. It can be divided into two procedures: off-line and on-line processing. For the off-line processing, several attribute classifiers are trained to detect the attributes of the gallery images and a distance metric based on attribute vectors is learned for the distance measure. For the on-line processing, a linear sparse reconstruction is conducted to complete the user-provided attributes. And then, the learnt distance metric is adopted for the completed attribute vectors.

The main contribution of this paper can be summarized as follows:

- A new issue which is very important for practical surveillance application is addressed, while the specific person retrieval via incomplete text description in camera network is seldom investigated.
- A linear sparse reconstruction [9] is introduced for incomplete user-provided attributes while the attribute incompleteness is seldom investigated in person retrieval. Metric learning is firstly conducted for attribute distance measure, while metric learning is usually utilized for visual descriptions.
- Extensive experiments conducted on two representative datasets, the VIPeR [5] and PRID2011 [6], have shown our superior performance.

2. THE APPROACH

The framework of our approach is shown in Fig.2. It can be divided into two main parts: *Off-line Processing* and *On-line Processing*. While the off-line processing mainly focus on training attribute classifiers and training distance metric, and the on-line processing introduces a linear sparse reconstruction to complete the original attribute vector, and then distance measure is conducted with the learned metric M to rank the results. The detail is discussed in the following.

2.1 Off-line Processing

The off-line processing pre-computes non-query images visual feature vectors and attribute vectors in order for the run-time query specific operations to be fast. Furthermore, to better computing their similarity, a pairwise-based metric learning framework is introduced to learn a distance metric.

Attribute SVM Training. A linear SVM is adopted to train the attributes classifiers for the reason that it is both fast to train and fast to test. The training data consists of the visual feature vectors extracted on the the training images, together with the pre-labeled attributes contributed by [3]. For each kind of attribute, a linear classifier is trained. We treat the ones contain the attribute as positive samples while the others as negative. The classifier is trained using the software package Vlfeat[1], which is optimized for linear classifiers and has complexity linear in the number of training samples. And the experiments have shown the superior classifying performance comparing with results reported in other papers.

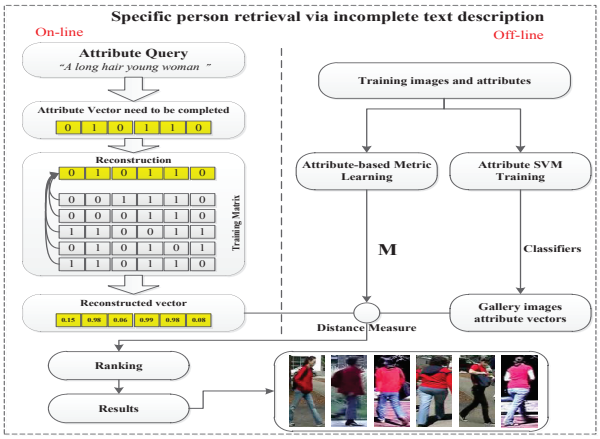


Figure 2: The framework of our retrieval system.

Gallery image attribute vectors. Based on the training classifiers, attribute vectors can be extracted for the gallery images. Note that the output value of the classifiers for each attribute is from minus infinity to plus infinity. But for better distance measure, we transform the original value to (0,1) by a logistic function as shown in

$$a = \frac{1}{1 + e^{-a_0}} \tag{1}$$

where the a_0 denotes the original output of the classifier, a is the transformed value that indicates the probability of owing this attribute. Specially, x is more closer to 1 indicates that it will more probably to own this attribute. "$a = 0$" expresses that the image do not have this attribute.

Attribute-based metric learning. For better measuring the attribute vector similarity, metric learning is conducted. Given a pair of samples x_i and x_j (x_i , $x_j \in R_d$), the Mahalanobis distance between them is:

$$d_M^2(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) \tag{2}$$

where $M \geq 0$ is a positive semi-definite matrix. Based on a statistical inference point of view, KISSME defines the Mahalanobis distance matrix M by

$$M = \Sigma_{y_{ij}=1}^{-1} - \Sigma_{y_{ij}=0}^{-1} \tag{3}$$

where

$$\Sigma_{y_{ij}=1} = \sum_{y_{ij}=1} (x_i - x_j)(x_i - x_j)^T \tag{4}$$

$$\Sigma_{y_{ij}=0} = \sum_{y_{ij}=0} (x_i - x_j)(x_i - x_j)^T \tag{5}$$

denote the covariance matrices for similar pairs $y_{ij} = 1$ and dissimilar pairs $y_{ij} = 0$ respectively. Then, M can be learned easily from the training samples. More details can be found in KISSME [7]. The learned matrix M is used for computing the distance between the attribute vectors.

[1]It is available in http://www.vlfeat.org/

2.2 On-line Processing

The on-line processing firstly formulate an original attribute vector by the text description, and then a linear sparse reconstruction is introduced to construct a completed vector. After that, with the per-trained distance matrix M in Section 2.1, the reconstructed attribute vector is adopted to query in the original gallery image set. And then, the final ranking results are achieved.

Attribute vector need to be completed. According to the user-provided text query, we formulate an original attribute vector $t_{1\times n}$, where n denotes the number of pre-defined attributes. Specially, the ones user provided is labeled as 1, the unlabeled ones is 0.

Reconstruction. Given a to-be-reconstructed vector $t_{1\times n}$, and the training attribute matrix $\widehat{T}_{m\times n}$, where m is the number of training images. Linear sparse reconstruction aims to reconstructing an attribute vector with the other images' attribute vectors $\widehat{T}$. Furthermore, based on the observation that images associated with an identical attribute tends to share more common semantic content and thus form a group [12]. For example, *"hair-long"* are more likely to be a *"female"*, we name it as *occurrence*. Moreover, another observation about *exclusiveness* is considered, for instance, *"male"* and *"female"*, which illustrates that the two attributes are unlikely to appeared at the same time. Therefore, a group sparse structure is introduced for the reconstruction weight. Here, we denote the reconstruction weighting vector as w, the ith group of reconstruction weight is defined as $g_i = \{w_{(i,1)}, w_{(i,2)}, \cdots, w_{(i,|g_i|)}\}$, where (i,j) is the index of the jth weight of the ith group in the weighting vector w. As can be seen that each attribute corresponds to a group of reconstruction weights, *i.e.*, the weights of images containing this attribute. Specially, higher weights for the *occurrence* and lower weights for *exclusiveness*. Since the images would be labeled with several attributes and thus the groups may be overlapped. And the reconstruction problem is formulated as:

$$J(w) = \min_{w} \left\| W(t - \widehat{T}w) \right\|_2^2 + \lambda \sum_{i=1}^{n} \|g_i\|_2 \qquad (6)$$

where $t_{1\times n}$ is the user-provided attribute vector, $\widehat{T}_{m\times n}$ is the dictionary matrix containing attribute vectors of the training images, $w_{m\times 1}$ is the objective reconstructing weighting vector, λ is a tuning factor for balancing the group sparsity. Specially, the group sparsity $\Sigma_{i=1}^{n}\|g_i\|_2$ is the combination of both $L1$ and $L2$ norms [12]. The $L1$ norm is used for emphasizing inter-group sparsity since only a few tags are associated with the target query image (*i.e.* inter-group sparsity). As for the reason that images in the same group are all supposed to contribute to the reconstruction if the corresponding attribute would be associated (*i.e.* intra-group smoothing), the $L2$ norm is adopted for the smoothing intra-group weights. Additionally, it is reasonable that the labelled attributes (*i.e.* non-zero entries) should be assigned higher weights. Therefore, a diagonal matrix for weighting the reconstruction residual of each entry in t is introduced, defined as $W_{i,i} = exp(t_i)$. It can be seen that the unlabelled ones (*i.e.* zero entries) are assigned with lower values.

With the linear sparse reconstruction, an optimal weighting vector w is obtained. Then the reconstructed vector $\widehat{t}$ is achieved by $\widehat{t} = \widehat{T}w$. Specifically, in our experiments the maximal value of $\widehat{t}$ is normalized as 1.

Table 1: Attribute detection accuracy on VIPeR.

Attributes	[8]	[3]	Ours	Attributes	[8]	[3]	Ours
shorts	0.74	0.56	**0.89**	male	0.68	**0.81**	0.66
jeans	0.73	**0.76**	0.73	backpack	0.52	0.67	**0.68**
vnecks	0.53	0.51	**0.81**	logo	0.58	0.51	**0.84**
sunglasses	0.60	0.52	**0.70**	stripes	0.47	0.52	**0.87**
longhair	0.55	**0.73**	0.63	skirt	0.76	0.64	**0.92**
sandals	0.58	0.50	**0.82**	**Average**	0.59	0.71	**0.76**

Ranking. By the reconstructed vector $\widehat{t}$ and the pre-processed gallery images attribute vectors, together with the learned distance matrix M, the distances between the query and the gallery images are obtained. According to the computed distances, the retrieval results are ranked.

3. EXPERIMENTS

3.1 Datasets and evaluation protocol

Datasets. Two publicly representative datasets, the VIPeR dataset [5] and the PRID2011 [6] dataset are adopted to verify our approach. We chose 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 with drastic appearance difference between most of the matched image pairs, and the images are normalized to 128×48 pixels. PRID dataset is a challenge dataset for person re-identification with more camera characteristics variation than VIPeR. Particularly, 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.

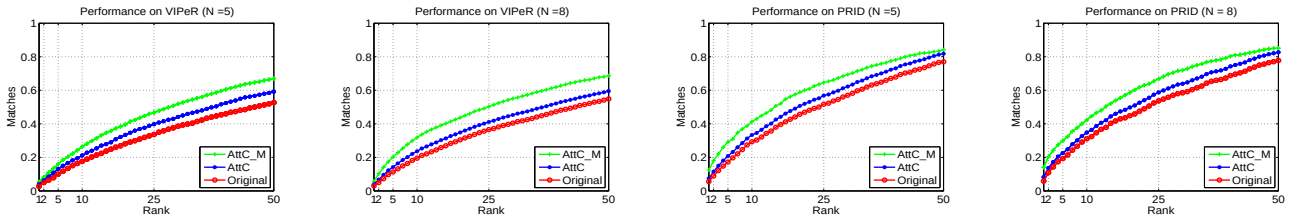
Features. The visual features are extracted for training the SVM classifiers and generate the attribute vector of the gallery images. Similar to [5], a combination feature descriptor consisting of color and texture features is conducted. Specifically, for each image, the RGB and HSV color histograms and LBP descriptor are extracted from overlapping blocks of size 16×16 and stride of 8×8.

Attributes. The attributes we adopted are derived from [3], *i.e.* PETA dataset. The dataset contains 61 kinds of attributes. Since some of the attributes are rarely appeared, we delete some of the attributes as done in [3], with only 35 pre-defined attributes are remained, such as *"personalFemale", "personLess30", "carryingbackpack".*

Evaluation protocol. For each dataset, we select half for training, while the retrieval performance is reported on the held out test portion. Firstly, we evaluate the performance of SVM classifiers based on the attribute detection accuracy. Moreover, we evaluate the retreival performance with the averaged cumulative match characteristic (CMC) curve [14] over 10 trials. Specially, we randomly select $5-8$ attributes of each query as the original text query. As recommended in [9], the tuning factor λ is set as $\lambda = 2$.

3.2 Attributes Detection

The attribute detection performance is reported with accuracy, some results on VIPeR dataset are shown in Table 1. With the SVM package based on [13], the average detection accuracy is about 76%, and more than a half attribute



(a) On VIPeR (N = 5) (b) On VIPeR (N = 8) (c) On PRID2011 (N = 5) (d) On PRID2011 (N = 8)

Figure 3: Retrieval results on VIPeR and PRID2011 dataset. N denotes the number of attributes of the original query. While "*Original* " is the original retrieval result. "*AttC* " indicates the attribute completion result and "*AttC_M* " represent the metric learning result after the attribute completion, respectively.

Table 2: Comparable results on VIPeR dataset.

Method	rank@1	5	10	25	50
W.AIR [8]	4.84	17.44	29.24	**50.60**	68.64
MLA [11]	5.06	19.24	29.06	48.06	68.00
AttRel [1]	*	12.51	23.34	40.70	*
AttC(N=8)	3.91	12.76	23.61	41.13	59.49
AttC_M(N=8)	**5.77**	**20.13**	**31.84**	50.29	**68.72**

detectors can achieve an accuracy above 70%, which is a little higher than the results reported in [1, 3, 8]. The attribute detection accuracy is on PRID2011 dataset is about 73%, while most of the attribute classifiers can achieve an accuracy over 65%.

3.3 Retrieval performance

The retrieval results are reported with CMC curves. Specially, to evaluate the attribute completion performance, we randomly select N attributes for each text query. The results are shown in Fig.3 and several conclusions can be drawn as follows. (1) Attribute completion is effective. The retrieval results are significantly improved with the attribute completion. Specially, the average improvement at rank@10 is about 17.2% and 15.7% for rank@25, respectively. (2) Metric learning for attribute vectors are still satisfying. While the metric learning is usually adopted for visual features and seldom conducted for attribute distance measure. As illustrated in our experiments shown in the figures, the improvements are noteworthy. (3) Query with more attributes is better. Based on the comparison that the retrieval results of $N = 8$ are better than $N = 5$, thus can be seen that with more attributes provided, the results is better.

Furthermore, the results compare favorably with the related attribute-based person re-identification works as shown in Table.2. Note that the related works are conducted based on the visual information of the query person while it is unused in our experiments.

Acknowledgement. The research was supported by National Nature Science Foundation of China (61303114, 61231015, 61170023), the Specialized Research Fund for the Doctoral Program of Higher Education (20130141120024), the Technology Research Project of Ministry of Public Security (2014JSYJA016), the China Postdoctoral Science Foundation funded project (2013M530350), the major Science and Technology Innovation Plan of Hubei Province (2013AAA020), the Key Technology R&D Program of Wuhan (2013030409020109), the Guangdong-Hongkong Key Domain Breakthrough Project of China (2012A090200007), and the Special Project on the Integration of Industry, Education and Research of Guangdong Province (2011B090400601).Nature Science Foundation of Hubei Province (2014CFB712).

4. CONCLUSION

In this paper, a new issue for specific person retrieval task in camera network is addressed. And a general framework to solve this issue is proposed. The original text query is transformed to an incomplete attribute vector and completed by a linear sparse reconstruction, and then a pairwise-based metric learning is adopted to refine the results. Extensive experiments show the superiority of our proposed approach. In the future work, better combination of visual information and attributes needs to be further investigated.

5. REFERENCES

[1] N. N. B, N. V. H, D. T. N, and et al. Attrel: An approach to person re-identification by exploiting attribute relationships. In *Multimedia Modeling (MMM)*, 2015.

[2] X. Cao, H. Zhang, X. Guo, S. Liu, and X. Chen. Image retrieval and ranking via consistently reconstructing multi-attribute queries. In *ECCV*, 2014.

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

[4] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

[5] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS workshop*, 2007.

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

[7] M. Kostinger, M. Hirzer, P. Wohlhart, and et al. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[8] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, 2012.

[9] Z. Lin, G. Ding, M. Hu, J. Wang, and et al. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *CVPR*, 2013.

[10] X. Liu, M. Song, Q. Zhao, D. Tao, and et al. Attribute-restricted latent topic model for person re-identification. In *PR*, 2012.

[11] L. R, H. T. M, and G. S. Towards person identification and re-identification with attributes. In *ECCV workshop*, 2012.

[12] Z. S, H. J, H. Y, and et al. Automatic image annotation using group sparsity. In *CVPR*, 2010.

[13] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. In *JMLR*, 2013.

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

[15] L. Wu, R. Jin, and A. Jain. Tag completion for image retrieval. In *PAMI*, 2013.

[16] M. Ye, J. Chen, Q. Leng, and et al. Copuled-view based ranking optimization for person re-identification. In *Multimedia Modeling (MMM)*, 2015.

[17] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.