Automatic Attribute Learning for Person Re-Identification

Yang Li\textsuperscript{1,2,*}, Huahu Xu\textsuperscript{1}, Minjie Bian\textsuperscript{1} and Junsheng Xiao\textsuperscript{1}

\textsuperscript{1} School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
\textsuperscript{2} School of Information Technology, Shanghai Jianqiao University, Shanghai 201306, China

*E-mail: liyang162@shu.edu.cn

Abstract. Attribute based person re-identification is more robust to image appearance changes than low-level visual feature based methods. However, manual person attribute annotation has low efficiency, poor scalability and poor discrimination. Therefore, a scalable automatic attribute learning method for person re-identification is proposed. Firstly, the mapping matrix of person attributes is designed and generated automatically according to the principles of discernibility, sharing and low redundancy. Then the visual features are extracted using convolutional neural network and attribute classifiers are trained to detect person attributes, and the attribute based representation of person is generated. Finally, the person re-identification is carried out by comparing the similarity of the attribute based representations. Extensive experiments have been carried out on two common datasets Market-1501 and Duke MTMC-reID with the comparison to the State-of-the-Arts methods, the method achieved the best performance, which proves the superiority and effectiveness of the method.

1. Introduction

Person re-identification needs to recognize the same person from different scenes shot by non-overlapping cameras. In large-scale camera networks, it is time-consuming, expensive, inaccurate and relying on manual person re-identification. Operators are prone to be distracted when manually identifying and matching persons in images from multiple cameras for a long time, and the recognition results are affected by the experience of individual operators and other human factors. In recent years, with the rapid development of public security video surveillance camera network, the research on automatic person re-identification has been widely valued.

The existing methods of person re-identification mainly focus on two aspects: one is to design the feature representation methods, the other is to design the learning model and optimize the parameters for person re-identification. Although the known methods have achieved good results, most of them are based on low-level visual features, they are sensitive to the changes of appearance (such as lighting, posture, occlusion, background complexity), have not enough robustness for the changes of visual conditions. Automatic person re-identification is still a technical problem to be solved.

Recent studies have proposed attribute based approaches [1,2], which use visual attributes (such as gender, hair length, clothing style, etc.) as features to calculate similarity rather than low-level features. Attribute, as a high-level abstraction of visual features, has obvious advantages in image classification, the method based on attributes has higher robustness to the changes of appearance and can effectively improve the performance of person re-identification.
At present, most attribute based methods usually need to manually select a set of attributes that can fully describe the involved image when designing attributes. The attribute set can be heuristic or selected from the knowledge base provided by domain experts. After the attribute set is determined, additional manpower is needed to label attributes. Because of the need of human participation and supervision, a large number of attributes are hindered. More importantly, although a set of manually defined attributes may be intuitive, they do not have enough discrimination for visual recognition tasks.

In this paper, an extensible automatic attribute design method is proposed, which can automatically learn a set of non-semantic attributes for person re-identification without human participation, and optimizes the attributes to make them have good discrimination between different persons and maintain robustness to the appearance changes of the same person. Through the learned attributes, a person can be represented as a group of specific attribute combinations. Person re-identification is to find the person with the most similar attribute combination with the interested person.

2. Related works

2.1. Person re-identification

Most person re-identification methods based on deep learning can be divided into two categories. The first is the deep metric learning method, which uses image pair [3] or triplet [4] as the input for Siamese model. The main idea of this method is to minimize the feature distance between the same person and maximize the feature distance between different persons. The second is to treat the task of person re-identification as an image classification problem [5,6]. In literature [5], it is suggested to combine multiple data sets for training, and improve feature learning through domain guided drop out. Literature [6] indicates that the classification model can achieve higher accuracy than the Siamese model. Recently, some semi-supervised and unsupervised methods [7,8] have been proposed to solve the problem of insufficient datasets for person re-identification. These methods achieve surprising performance with little or no annotation.

2.2. Attribute learning

Literature [1] uses low-level features and support vector machine (SVM) to train attribute detector, and integrates attributes through several metric learning methods. Literature [2] deals with person re-identification in a systematic way using low-level features learned from attributes and camera correlation. In recent years, some deep learning methods have been proposed. Literature [9] first trains the network on the independent dataset with attribute labels, then fine tunes the network on the target dataset with only person ID labels, and finally combines the prediction attribute labels of the target dataset with the independent dataset for the last round of fine tuning. In literature [10], an unsupervised person re-identification method is proposed, in which the source domain knowledge is learned from the source data with attribute annotation, and the knowledge is transferred to the target dataset without annotation through the cross domain joint attribute identity transfer learning.

In order to improve the efficiency and performance of person re-identification, this paper uses the non-semantic attributes from automatic learning to represent persons, and improves the robustness and discrimination of large appearance changes caused by different viewpoints, lighting changes, deformation, etc.

3. Automatic attribute learning method for person re-identification

3.1. Selection of attribute generation method

As a higher-level abstraction of visual features, attributes are more robust to changes in appearance (such as viewpoints, lighting, occlusion, background, etc.). Attributes can be divided into two types: manually defined attributes with semantics [9] and automatically learned attributes without semantics [11]. It takes a lot of manpower and domain expertise to obtain manually defined attributes, which
makes it difficult to extend to a large number of attributes. In addition, manually selected attributes are not necessarily machine detectable and may not very useful for the tasks. The attributes of automatic learning do not need manual annotation, have good scalability, and have the ability of task optimization. For some tasks, such as zero shot learning and image retrieval based on text query, it is necessary to use human understandable semantic attributes. However, for person re-identification based on image query, it is not necessary to use semantic attributes. This paper uses the non-semantic attributes with automatic learning for person re-identification.

3.2. Automatic attribute learning framework for person re-identification

Given a set of training images of person, our goal is to learn the attribute representation of persons, and use them for person re-identification. Specifically, given \( n \) different persons, if \( k \)-dimension attribute is used to represent persons, then the attribute matrix of persons can be expressed as \( A \in \mathbb{R}^{n \times k} \), the row of matrix represents \( n \) different persons, and the column of matrix represents \( k \) attributes.

The framework of attribute based person re-identification algorithm is shown as Figure 1, it has two main parts: attribute based person feature extraction and person similarity estimation. For the first part, we need to design the attribute mapping matrix, then use convolutional neural network (CNN) to extract visual features of the image and train \( k \) attribute classifiers to detect the attributes in the image, and use \( k \)-dimension attribute to detect the confidence score to form the attribute based feature representation of the person image. In the second part, the image of probe person and images from the gallery are represented by \( k \)-dimension attribute detection scores. By comparing the distance in \( k \)-dimension attribute feature space to re-identify the person. This paper focus on the first part of attribute matrix design and the person feature extraction based on attributes.

![Figure 1. Person re-identification framework based on attributes.](image)

3.3. Person attribute mapping matrix design

Before extracting attribute based person features, it is necessary to build a person attribute matrix. In this paper, the mapping matrix of person attributes is generated automatically. In order to make the generated attribute matrix efficient for person re-identification, three principles should be followed in the design of attribute mapping: distinguishable, shareable and low redundancy. Because the goal of person re-identification is to identify different persons and find out the persons with high similarity, the attribute should be able to distinguish between the persons to be identified. On the other hand, because these attributes will be used for identification of unseen persons, the learned attributes need to be able to generalize to unseen persons. For this reason, visually similar persons need to share some attributes. In addition, in order to make full use of \( k \)-dimension attributes, it is necessary to have low redundancy between these attributes. Considering the above factors, in order to optimize the design of person attribute matrix, the objective function can be expressed as follows:

\[
\max_{A} f_1(A) + x f_2(A) + y f_3(A)
\]

The \( f_1(A) \), \( f_2(A) \) and \( f_3(A) \) in formula (1) are defined as follows:

\[
f_1(A) = \sum_{i,j} \|A_i - A_j\|_2^2
\]
\[ f_2(A) = -\sum_{i,j} S_{ij} \| A_{ij} - A_{ij} \|_2 \]

\[ f_3(A) = -\| A^T A - I \|_F \]

where \( A_i \) represents the \( i \)-th row in attribute matrix \( A \), which is the attribute feature of the \( i \)-th person. \( f_1(A) \) ensures the distinguishability of persons. In order to make the algorithm more effective, \( f_1(A) \) is set as the sum of all the distances between each two rows in matrix \( A \), encouraging that every two different persons in the attribute space are separable. \( S \) in \( f_3(A) \) is the visual proximity matrix, where \( S_{ij} \) represents the visual proximity between person \( i \) and person \( j \), and the proximity between two persons is calculated as the average proximity of attributes between two person images. \( f_3(A) \) encourages using similar attribute representations between visually similar persons, resulting in shareable attributes. It means if two persons are visually similar, they are expected to share more attributes. \( f_3(A) \) penalizes redundancy between attributes. \( x \) and \( y \) are two parameters of the objective function. The larger \( x \) encourages more attribute sharing among visually similar persons, while the larger \( y \) will form more punishment for the redundancy of the learned attributes.

Person attribute matrix \( A \) is a step-by-step generation and optimization process, a column of \( A \) obtained in each step is an attribute, the process is briefly described below.

The objective function formula (1) can be rewritten as:

\[ \max_A \text{Tr}\left( A^T P A \right) - y \| A^T A - I \|_F \]

where \( P = Q - xL \), \( Q \) is an \( n \times n \) matrix, the diagonal elements of the matrix are \( n-1 \), and the non-diagonal elements are \(-1\). \( L \) is the Laplace representation of \( S \). \( A \) is initialized as an empty matrix. Following formula can be used to optimize and add a column of \( A \) by incremental learning:

\[ \max_{a} a^T Ra \text{ s.t. } a^T a = 1 \]

where \( R = P - 2yAA^T \), \( a \) is the new column of attributes add to the attribute matrix \( A \), and the optimal \( a \) is the eigenvector of \( R \) with the largest eigenvalue. Matrix \( A \) is updated to \( A = [A, a] \) at each step.

The algorithm of automatic generation of person attributes in this paper is an efficient and scalable algorithm, which greedily searches for new non-redundant attributes that meet the expected values every iteration.

3.4. Person attribute detection

After getting the attribute mapping matrix \( A \), the next step is to learn the attribute detector. Support vector machine (SVM) classifier is used for attribute detection in this paper. The low-level features of person image are extracted by fine tuning the VGG16 convolutional neural network model pre-trained on ImageNet, and then it is sent to SVM classifier. By using LIBSVM and RBF kernel, the classifier is trained to detect whether an attribute appears in the image. For each attribute, the parameters \( C \) and \( \gamma \) of SVM are obtained by cross validation. The output of attribute classifier is used to predict the confidence score of attribute appearance in image. When training the detector of the \( j \)-th attribute, the person image with \( A_j > 0 \) is taken as a positive sample and the rest as negative samples.

3.5. Person attribute feature representation

For a person image, attribute classifier is used to detect all attributes in the image. Then combine the confidence scores of each attribute to generate the final attribute based feature vector of the person image, that is, the attribute feature representation of the person image, which can be specifically expressed as:

\[ \overline{f_I} = \{ S_{A_1}, S_{A_2}, \ldots, S_{A_k} \} \]

where \( \overline{f_I} \) is the feature vector of person image \( I \), \( S_{A_i} \) is the confidence score representing the occurrence of attribute \( A_i \) in image \( I \).
The attribute feature representation is helpful to improve the discrimination of different persons, because the attribute matrix of persons is optimized for this purpose. In addition, the representation of person’s attribute features has strong robustness to the appearance change of person image.

3.6. Person similarity estimation
After getting the attribute based feature vector of each person image, the similarity between images can be calculated to determine which images belong to the same person. This paper uses L2 distance to measure the similarity between two images. The similarity is calculated as follows:

\[ sim(f_1, f_2) = \sqrt{(S_{A_1} - S_{A_1}^2)^2 + \cdots + (S_{A_k} - S_{A_k}^2)^2} \]

(8)

In this paper, the standard L2 distance is used for similarity calculation, because the main goal of this paper is to evaluate the effectiveness of the designed automatic attribute learning method for person re-identification, rather than to evaluate the matching method.

4. Experimental results and analysis

4.1. Dataset and evaluation metrics
The experiments are carried on two large-scale person re-identification datasets Market-1501 [12] and DukeMTMC-reID [13].

Market-1501 is one of the largest and most challenging person re-identification datasets. The dataset contains 19732 images of 751 person IDs for training and 13328 images of 750 person IDs for testing. Each image has 27 attributes labelled.

The DukeMTMC-reID dataset is a subset of the DukeMTMC dataset and is designed for person re-identification. The dataset is divided into 16522 training images of 702 person IDs and 19889 test images of 702 person IDs. Each image has 23 attributes labelled.

Evaluation metrics. For person re-identification task, cumulative matching feature (CMC) curve and mean accuracy (mAP) are used for evaluation. CMC is used for the case that only one query person image is included in the tested person gallery. Person images are sorted from small to large according to the distance between the query image and the images in the gallery. CMC uses Rank-k to represent the index of sorting hit rate. Rank-1 is the first hit rate, that is, whether the image ranked first. CMC is not suitable for the situation that there are multiple person images in the gallery are the same person as the query, mAP is more suitable for this situation, mAP reflects the extent to which all the correct images of the queried person in the gallery are in front of the result queue.

4.2. Parameter setting
The influence of two parameters \( x \) and \( y \) of formula (1) in the automatic attribute learning algorithm on the performance of person re-identification is experimented and evaluated.

First study the influence of \( x \) by fixing \( y \). In an extreme case, \( x \) is set to 0, which means that there is no attribute sharing between the training person images. When \( x \) is 0, recognition performance is much worse than when \( x \) is 1 to 5, especially when the number of attributes is low. This shows that when there is no shared attribute guidance, the learned person attributes cannot be generalized to new persons in the test dataset. As long as attribute sharing is enabled, the re-identification performance is robust to the value of \( x \).

Then study the influence of \( y \) by fixing \( x \). When \( y \) is small, it means that there is a large redundancy in the automatic learning attributes, the re-identification performance is low, but when \( y \) is large enough, the re-identification performance will be improved and stable.

The above research shows the importance of forced attribute sharing and low redundancy in the learning process, and the robustness of automatic attribute learning algorithm to \( x \) and \( y \) values for person re-identification performance. According to the above evaluations, in the other experiments of this paper, \( x \) and \( y \) are set to 2 and 7 respectively to achieve the best experimental results.
4.3. Comparison with state-of-the-arts methods

The performance of our method is compared with other State-of-the-Arts (SOTA) methods using CMC and mAP as evaluation metrics on Market-1501 and DukeMTMC-reID datasets, the results are shown in Table 1. The methods are divided into three categories: traditional methods based on manual feature and distance metric learning, including Bow+kissMe [12], CRAFT [14], and LOMO+XQAD [15]; methods based on deep learning, including SVDNet [16], PAN [17], MultiScale [18] and HACNN [19]; methods based on attributes, including ACRN [20], APR [21] and CA3net [22].

Table 1. Performance comparison of methods on Market-1501 and DukeMTMC-reID.

| Method          | Market-1501 Rank-1(%) | mAP(%) | DukeMTMC-reID Rank-1(%) | mAP(%) |
|-----------------|------------------------|--------|-------------------------|--------|
| Bow+kissMe [12] | 44.4                   | 20.8   | 25.1                    | 12.2   |
| CRAFT [14]      | 68.7                   | 42.3   | -                       | -      |
| LOMO+XQAD [15]  | -                      | -      | 30.8                    | 17.0   |
| SVDNet [16]     | 82.3                   | 62.1   | 76.7                    | 56.8   |
| PAN [17]        | 82.8                   | 63.4   | 71.6                    | 45.0   |
| MultiScale [18] | 88.9                   | 73.1   | 79.2                    | 60.6   |
| HACNN [19]      | 91.2                   | 75.7   | 80.5                    | 63.8   |
| ACRN [20]       | 83.6                   | 62.6   | 72.6                    | 52.0   |
| APR [21]        | 84.3                   | 64.7   | 70.7                    | 51.9   |
| CA3net [22]     | 93.2                   | 80.0   | 84.6                    | 70.2   |
| Ours            | 94.1                   | 81.2   | 85.7                    | 71.9   |

Compared with other SOTA methods, our method achieves the best performance on both datasets. The experimental results show the advantages of the proposed automatic attribute learning method in person re-identification. This is mainly because on the one hand, compared with methods based on low-level visual features (including traditional methods based on manual features and methods based on deep learning), attributes are abstractions of low-level visual features, belonging to high-level features, which have higher robustness to appearance changes (such as lighting, viewpoint, posture, etc.), and can effectively improve the performance of person re-identification. On the other hand, although the manually labelled person attributes have certain semantics and are relatively intuitive, they are not necessarily suitable for automatic computer vision recognition. Moreover, due to the limited conditions, there are only dozens of person attributes manually labelled, which is difficult to provide enough discrimination ability for person re-identification. In this paper, the attribute automatic learning method optimizes when the attribute is generated automatically to meet the discrimination requirements of person re-identification. Automatic attribute learning has good scalability, which can easily generate more attribute classes according to task needs. Therefore, compared with the method based on manual annotation of attributes, automatic attribute learning has better task pertinence and scalability, has higher discrimination, and can effectively improve the performance of person re-identification. This also shows the effectiveness and superiority of our method.

5. Conclusion

In this paper, based on the analysis of the shortcomings of manual attribute labelling, an automatic attribute learning method for person re-identification is proposed. This method does not need human participation, automatically generates person attribute mapping matrix, performs attribute detection and generates person attribute features, with high efficiency and good scalability, and optimized for the person re-identification task, with strong discrimination and robustness. Experimental results show that the method has the best person re-identification performance, which proves the effectiveness and superiority of the method. In the future work, we will further study the portability and scalability of automatic attribute learning, and apply the method to cross domain person re-identification.
References

[1] Hospedales T, Layne R, Gong S. Re-id: Hunting attributes in the wild[C]//British Machine Vision Conference. 2014.

[2] Su C, Zhang S, Yang F, et al. Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping[J]. Pattern Recognition, 2017, 66: 4-15.

[3] Xu J, Zhao R, Zhu F, et al. Attention-aware compositional network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2119-2128.

[4] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1335-1344.

[5] Xiao T, Li H, Ouyang W, et al. Learning deep feature representations with domain guided dropout for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1249-1258.

[6] Zheng L, Zhang H, Sun S, et al. Person re-identification in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1367-1376.

[7] Wu Y, Lin Y, Dong X, et al. Progressive learning for person re-identification with one example[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2872-2881.

[8] Lin Y, Dong X, Zheng L, et al. A bottom-up clustering approach to unsupervised person re-identification[C]// Proceedings of AAAI Conference on Artificial Intelligence. 2019, 2: 1-8.

[9] Su C, Zhang S, Xing J, et al. Multi-type attributes driven multi-camera person re-identification[J]. Pattern Recognition, 2018, 75: 77-89.

[10] Wang J, Zhu X, Gong S, et al. Transferable joint attribute-identity deep learning for unsupervised person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2275-2284.

[11] Sharmanska V, Quadrianto N, Lnapert C H. Augmented attribute representations[C]//European Conference on Computer Vision. 2012: 242-255.

[12] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116-1124.

[13] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3754-3762.

[14] Chen Y C, Zhu X, Zheng W S, et al. Person re-identification by camera correlation aware feature augmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(2): 392-408.

[15] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2197-2206.

[16] Sun Y, Zheng L, Deng W, et al. SVDnet for pedestrian retrieval[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3800-3808.

[17] Zheng Z, Zheng L, Yang Y. Pedestrian alignment network for large-scale person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018: 99.

[18] Chen Y, Zhu X, Gong S. Person re-identification by deep learning multi-scale representations[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2590-2600.

[19] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2285-2294.

[20] Schumann A, Stiefelhagen R. Person re-identification by deep learning attribute-complementary information[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 20-28.

[21] Lin Y, Zheng L, Zheng Z, et al. Improving person re-identification by attribute and identity learning[J]. Pattern Recognition, 2019.

[22] Liu J, Zha Z J, Xie H, et al. CA3Net: Contextual-Attentional Attribute- Appearance Network for Person Re-Identification[C]//2018 ACM Multimedia Conference on Multimedia Conference. 2018: 737-745.