Attributes-based person re-identification via CNNs with coupled clusters loss

SUN Rui¹.²,* HUANG Qiheng¹.², FANG Wei¹.², and ZHANG Xudong¹.²

1. Key Laboratory of Knowledge Engineering with Big Data (Ministry of Education), Hefei University of Technology, Hefei 230601, China; 2. School of Computer and Information, Hefei University of Technology, Hefei 230601, China

Abstract: Person re-identification (re-id) involves matching a person across nonoverlapping views, with different poses, illuminations and conditions. Visual attributes are understandable semantic information to help improve the issues including illumination changes, viewpoint variations and occlusions. This paper proposes an end-to-end framework of deep learning for attribute-based person re-id. In the feature representation stage of framework, the improved convolutional neural network (CNN) model is designed to leverage the information contained in automatically detected attributes and learned low-dimensional CNN features. Moreover, an attribute classifier is trained on separate data and includes its responses into the training process of our person re-id model. The coupled clusters loss function is used in the training stage of the framework, which enhances the discriminability of both types of features. The combined features are mapped into the Euclidean space. The L2 distance can be used to calculate the distance between any two pedestrians to determine whether they are the same. Extensive experiments validate the superiority and advantages of our proposed framework over state-of-the-art competitors on contemporary challenging person re-id datasets.

Keywords: person re-identification (re-id), convolutions neural network (CNN), attributes, coupled clusters loss (CCL).

DOI: 10.21629/JSEE.2020.01.06

1. Introduction

Person re-identification (re-id) is about matching the problem of re-identifying the people across various camera views at different times and locations. In real scenes, person re-id has very important applications including public safety and intelligent video surveillance. Nowadays it is attracting great interest from researchers in the fields of computer vision and pattern recognition [1 – 3]. The existing person re-id algorithm can be roughly categorized into two kinds: learning a robust feature representation [4 – 6] and a discriminative distance measure [7 – 10].

In recent years, some research works [11 – 13] on the subject has employed pedestrian attributes to improve the capacity of feature representation. The same person’s appearance variation and different persons’ appearance similarity may emerge during the feature representation stage of person re-id, resulting in difficulty in obtaining distinguishing information when representing pedestrians. Therefore, applying the information of pedestrian attributes can describe a person’s appearance more completely. In terms of person re-id, pedestrian attributes are potential and understandable high-level feature descriptors that are different from general image features such as colors and texture. Moreover, attributes are used as supplements to pedestrian identities. The knowledge of person attributes can help person re-id overcome the change of viewpoints to the robustness of the inner visual appearance. However, a still crucial challenge is how to fully use of attribute information for identifying a person.

To address the above issues, the improved convolutional neural network (CNN) is exploited to extract attribute features and low-level appearance features, and this paper uses the coupled clusters loss (CCL) to jointly train the features. Nowadays, CNNs have proved to be a very successful tool for person re-id. In [14], Li et al. proposed a filter pairing neural network (FPNN) to jointly handle misalignment, photometric and geometric transforms, occlusions and background clutter. All the key components are jointly optimized to maximize the strength of each component when cooperating with others. In contrast to existing works that use handcrafted features, the method automatically learns features optimal for the re-id task from data. In [15], Cheng et al. presented a multi-channel parts-based CNN model under the triplet framework for person re-id. Specifically, the CNN model consists of multiple channels to jointly learn both the global full-body and local body parts features of the input persons. In [16], Varior et al.
proposed a gating function to selectively emphasize such fine common local patterns by comparing the mid-level features across pairs of images. This produces flexible representations for the same image according to the images they are paired with. However, these algorithms mainly focus on describing the appearance features of a person, lacking the representation of semantic information. Therefore, this paper will demonstrate that pedestrian attributes can be used as additional, effective information to help solve the problem of pedestrian re-id. Compared with previous methods, we train the attribute features classifier on the independent pedestrian attribute dataset and incorporate the attribute features classifier into the training process of the pedestrian recognition model. In addition, the attribute feature extraction is integrated into the improved CNN learning framework for person re-id, and the network learns appearance features and attribute features at the same time.

The content of the rest work is organized as follows. In Section 2, we will discuss the related work about person re-id, visual attributes and deep learning. We will elaborate the proposed model in Section 3. Experiments about our attribute-based re-id method are outlined in Section 4. Section 5 is devoted to concluding remarks.

2. Related work

The person re-id system is performed in a network consisting of multiple camera sensors. The work platform used for person re-id is shown in Fig. 1. Fig. 1 illustrates the fact that person re-id can deal with the pedestrian matching and pedestrian matching score ranking problems in the non-overlapping camera view, but there are still many challenges including illumination changes, viewpoint variations and occlusions, which result in appearance inconstancy of the same person in two images.

Visual attributes can help us enhance the performance of person re-id. Fig. 2 describes a number of pedestrian attributes such as the gender, age and cloth colors or styles that can help distinguish examples of people with very similar visual appearance.

Because this work is mainly related to the topics of person re-id, visual attributes and deep learning, we briefly review the most recent methods which rely on the three fields including the following.

2.1 Person re-id

The early research work mainly focus on how to construct effective pedestrian characteristics. Large batch of different features have been proposed for person re-id, such as colors, textures, edges, shapes, global features, regional features and block-based features. In [4], the spatio-temporal local features were extracted from multiple consecutive frames for the region-based multiple segmented image. In [5], the symmetry-driven accumulation of local features was proposed, by utilizing the symmetry features in the pedestrian image to handle the view changes. In [6], the method used the color distribution field to extract color information, and used the differential excitation in the Weber local descriptor (WLD) to extract the texture information in the image, and used the direction in the WLD to extract the gradient direction.

Based on the features of extraction, the simple feature matching or metric learning method can get the moderate results on this basis. The basic idea of measurement learning is to find a mapping function that makes the features of the same person more similar to the features vectors of different pedestrians. These measures include KISSME [7], local fisher discriminant analysis (LFDA) [8], marginal fisher analysis (MFA) and locally adaptive decision functions (LADF) [9], and deep transfer metric learning (DTML) [10].

2.2 Visual attributes

Visual attributes are different from general image features such as colors, patterns and shapes, and help obtain a mid-level property for object recognition [17]. Visual attributes can also be used for zero-shot learning [18] and behavior recognition [19]. If an attribute describes a holistic pro-
perty in the image, e.g., ‘beautiful girl’, we call it a global attribute. Usually, global attributes do not refer to specific object parts or locations [20]. Local attributes can describe a part or several locations of the object, e.g., ‘pink pants’. The attributes learning method focuses on representing data instances by projecting them onto a set of foundations defined by domain-specific applications, which are semantically meaningful to humans and are preferred for subjective interpretation.

There are some methods proposed for attribute learning. Zhang et al. proposed to automatically learn attributes from an arbitrary set of image pairs [21]. Chen et al. proposed an attribute learning method to automatically acquire training images from publicly available images in an unsupervised manner [22]. To reduce the consumption of computation and storage resources, Lin et al. proposed a learning framework to detect various facial attributes such as gender and age to approximate more facial attributes [23]. In the field of video surveillance, the semantic attributes have begun to be applied in the research of attribute-based person search, which aims to retrieve person matching the description of language attributes from the camera network. However, these studies so far have only used relatively simple data, for example a small part of facial attributes. Most of the existing attribute learning methods learn linear functions to map hand-crafted features to the corresponding attributes. Different from these methods, the proposed method can learn image features and more effective nonlinear functions for attributes in a unified deep learning framework.

2.3 Deep learning

Deep learning is an effective feature learning method using deep neural networks. Its goal is to extract the high-level abstractions of visual data with multiple nonlinear transformations. There is a great number of deep models, e.g., deep belief networks, deep Boltzmann machines, stacked auto-encoders and CNNs. Among the deep learning models, CNN shows outstanding performance, specifically in image classification and recognition applications. A typical CNN consists of three parts [24].

(i) Convolution operator, which extracts local features through a set of learnable filters.

(ii) Activation function, which makes the network learn nonlinear transformation of the input data.

(iii) Pooling operator, which gathers features from adjacent locations and provides local spatial invariance to the extracted features.

Inspired by the encouraging performance of deep learning, some researchers adopt deep learning to learn visual features and distance metrics for person re-id. Yi et al. [25] proposed a siamese CNN for distance metric learning. Ahmed et al. [26] designed a deep neural network to classify a person from two cameras. Ding et al. [27] presented a scalable distance learning structure based on the deep neural network with the triplet loss. Liu et al. [28] proposed an end-to-end comparative attention network, which learns to selectively focus on parts of pairs of person images and adaptively compares their appearance. The above mentioned methods are designed specifically for certain datasets and are dependent on their camera settings. The deep network is used to acquire general camera-independent mid-level features in this paper. As a result, our method shows better robustness to handle person re-id tasks on datasets containing different types of images.

3. Proposed method

We use the deep learning method to include these attribute information into our proposed model. That is an end-to-end network framework designed specifically for person re-id. The basic idea of this model is to minimize the distance between the same pedestrian image and maximize the distance between different pedestrian images. Thus, we introduce the CCL function into the network framework. As shown in Fig. 3, we need to input two image sets into the model: one is a positive sample set (the identity of the pedestrian in the image is the same person) and the other is a negative sample set (the identity of the pedestrian in the image is not the same person). Loss of coupled clusters is to pull the positive sample image closer and push the negative sample image farther. The recognition performance of our method is obviously better than many of the most advanced methods, including traditional methods and methods based on deep networks, on challenging iLIDS-VID, PRID2011 datasets.

![Fig. 3 Framework of our model for person re-id](image-url)

3.1 Multiview attribute detection

The attributes of the pedestrian can be divided into two categories. Some of the attributes such as trousers or backpack are local, so only in certain areas of the image is visible, and other attributes, such as age or gender, are global.
attributes, which cannot be positioned to a single image area. Based on these, a CNN is used to extract multiview attribute features, which aggregates local and global image information. Fig. 4 shows a structure of attribute detection. We set up our network on a trained GoogleNet that is pre-trained in the dataset of ImageNet [29]. The GoogleNet is used to get global features on the image. The output of the initial five layers connecting to a fully connected layer with a size of 600 generates the global information of the pedestrian. Our method integrates local features associated with specific attributes into the GoogleNet architecture, so we split the layer about conv1 7x7 of the GoogleNet into three equal landscape areas. Every area expresses a partial view of the pedestrian image. According to the spatial locations, we split pedestrian into three views: the head-shoulder, the upper body and the lower body. For any of these views, we would train a sub-network (local view network) that can capture specific region features.

![Fig. 4 Structure of attribute detection](image)

The local view network (LVN) consists of double convolutional layers, followed by double residual blocks, each of which performs the function of pooling. The last section in the LVN is a layer which is fully connected with a size of 200. GoogleNet and LVN both execute batch standardization. The local and global attributes captured by the LVN and the GoogleNet are finally combined with a layer with the size of 2 048. The combination of the multiview attribute information is then used for attribute classification. However, this method of using multiple tags for classification is susceptible to imbalances in training data sets. Some attribute features appear more frequently than other attribute features, and this effect cannot be solved by sample sampling. Because some attributes are common, balancing the frequency of occurrence of one attribute may affect the frequency of another attribute. To solve this problem, we assign different weights to different attribute features in the loss function.

\[
L_{wce} = \sum_{i=1}^{L} \frac{1}{2w_i}p_i \cdot \log(q_i) + \frac{1}{2(1-w_i)}(1-p_i) \cdot \log(1-q_i)
\]

Among them, \(w_i\) is the weight of the attribute \(i\), and the weight reflects the ratio of occurrences in the training data set (the ratio of positive and negative samples), \(p_i\) is the prediction of attribute characteristics and \(q_i\) is the corresponding label.

We train our network on the dataset of PEdesTrain Attribute (PETA) [30]. For the attribute classification of GoogleNet using a fine-tuning strategy, we would set the initial value of the learning rate to 10-4; meanwhile the decay rate of the learning rate is set to 0.1. The LVN weight parameter is initialized randomly and the initial value of the learning rate is set to a relatively large value.

### 3.2 Person re-id based on attribute

When people identify a pedestrian target, they usually use basic information, i.e., color, texture, etc. When there are some factors that cause these low-dimensional features to fail, they can be distinguished by using attribute characteristics. For example, gender, age, hat, trousers, and slippers can all be used as attribute features to effectively identify targets. Similar to the human visual processing process, an end-to-end network model combines attribute features and low-dimensional features extracted from the CNN to extract more effective pedestrian features. An integrated person re-id network model is formed for improving pedestrian re-id performance. Fig. 5 shows the network...
model. The network extracts the attribute features and low-dimensional features with CCL.

![Person re-id network model](Fig. 5)

### 3.2.1 CCL

In general, the person re-id method uses a triad training sample to train the network. In the standard triple network, the input is an image triple \( \{< x^a, x^p, x^n >\} \), where \( x^a \) and \( x^b \) are the same person, and \( x^a \) and \( x^n \) are not the same person. Assume \( f(x) \) represents the network feature representation of the image \( x \). For image triples \( < x^a, x^p, x^n > \), their ideal characteristics indicate that the following formula should be satisfied:

\[
\|f(x^a) - f(x^p)\| + \alpha \leq \|f(x^a) - f(x^n)\| \tag{2}
\]

or equally

\[
\|f(x^a) - f(x^p)\|^2 + \alpha \leq \|f(x^a) - f(x^n)\|^2 \tag{3}
\]

where \( \alpha \) is a predefined constant, indicating the minimum interval between different classes. Fig. 6 shows (2) in a more obvious way. Therefore, the target loss function can be expressed as

\[
L = \sum_{1}^{N} \max\{\|f(x^a) - f(x^p)\|_2^2 + \alpha - \|f(x^a) - f(x^n)\|_2^2, 0\} \tag{4}
\]

Fig. 6 Triplet loss

However, when dealing with randomly selected triples, there are some special cases where the triples loss can be misjudged. Given three samples, the identities of the two of them are the same, the other is a different sample, and there are two different ways to construct the triples as input data for the network. Fig. 6 explains them. In the case of the left, because the distance between the same class is greater than that between different classes, it can very simply detect unusual distance relationships.

The right situation is a little different. Because of the distance between the anchor and the negative point, the triplet loss is 0. Therefore, the network will ignore the triples during the backward propagation.

Also, since the triple loss function of back propagation is actually to pull positive points to the anchor point, this will make the negative point move to the anchor point in the opposite direction, so the loss function is very sensitive to the choice of anchor. That means that improper anchor points could cause a lot of interference during the training period, leading to the slow speed of convergence. We need to use a better pedestrian image triple to correct it.

In order to stabilize the training stage and speed up the convergence speed, this paper uses a new loss function to replace the original triple loss function: CCL. We also use a deep convolutional network to extract the effective features of each pedestrian image. The difference between this method and the previous methods is that the original image group is replaced by two different image sets: positive and negative sets. The first set \( X^p = \{x^p_1, \ldots, x^p_{NP}\} \) includes \( NP \) images of the same person and the second one \( X^n = \{x^n_1, \ldots, x^n_{Nn}\} \) includes \( Nn \) images of different persons. In the Euclidean space, pedestrian samples with the same identity tag should be centered around the common center point. Therefore, the data in the positive sample set should form a cluster, and the samples in the negative sample set should be relatively far from the positive sample cluster. Fig. 7 illustrates the ideal situation.

Fig. 7 Triplets in two different situations

We first suppose the central point as the average of all positive samples.

\[
\mathbf{c}^p = \frac{1}{NP} \sum_{i}^{NP} f(x^p_i) \tag{5}
\]

The corresponding distance relation is expressed as

\[
\|f(x^a_i) - \mathbf{c}^p\|^2 + \alpha \leq \|f(x^p_j) - \mathbf{c}^p\|^2, \quad \forall 1 \leq i \leq NP; 1 \leq j \leq Nn. \tag{6}
\]

The CCL function is defined as

\[
L(W, X^p, X^n) = \sum_{i}^{NP} \max\{0, \|f(x^p_i) - \mathbf{c}^p\|_2^2 + \alpha - \|f(x^n_j) - \mathbf{c}^p\|_2^2\} \tag{7}
\]
where $x^*_n$ is negative and the closest to the center point. If $\|f(x^*_n) - c^p\|^2 + \alpha - \|f(x^*_n - c^p)\|^2 \leq 0$, the partial derivative of the positive and negative sample sets is 0. Otherwise, the partial derivative of the positive sample set is
\[
\frac{\partial L}{\partial f(x^p_i)} = f(x^p_i) - c^p.
\] (8)

The partial derivative of the nearest negative sample from the center point is
\[
\frac{\partial L}{\partial f(x^n_i)} = c^p - f(x^*_n).
\] (9)

The key idea behind (6) is exactly the same as the triple restrict in (2) that the distance between the same person should be less than the distance between different persons. But we express it to be a completely different way. First, the distance here changes to the distance between the sample and the cluster center, instead of randomly selecting a sample as the center point. Second, the definition of the CCL function is based on multiple samples, not randomly selecting three samples. The first requirement makes the distance we get and the direction in which the sample moves more reliable than the three samples we initially used. The second requirement is that the CCL function ensures that all pedestrian positive samples not too close to the center point of the sample are continuously zoomed in. Fig. 8 shows the principle of CCL.

3.2.2 Attribute-complementary re-id network

The positive and negative sample sets are used to train the network as shown in Fig. 9. Equation (7) is the loss function of our entire network, where $f$ represents the output of the entire network of this article, and $\alpha$ represents the minimum margin. The effect of the loss function is that the distance between similar samples is closer than the distance between different types of samples. For each input pedestrian image, there are altogether two branches to process it: one of them is used to identify the attribute features, and the other is for the extraction of low-level features based on CNN, as shown in Fig. 9.
The proposed CNN consists of three repeated convolutions, batch normalization and pooling layers, and uses the ReLU activation function in the network. The first size in the three convolution kernels is $5 \times 5$, the size of the other two convolution kernels is $3 \times 3$, the kernel size of the largest pooled layer is $2 \times 2$, and the number of channels in the convolution and connection layer is 32. Then, the two-level one-dimensional horizontal convolutions with a size of $3 \times 1$ instead of zero padding is used to reduce the feature map to a single column. The number of parameters in these network layers is relatively small and it is possible to model the displacements existing in horizontal stripes effectively. Then, the final CNN module output vector represents a feature of each horizontal stripe. In the final convolution, the number of channels increases to 150, followed by a 2 048-dimensional full connection layer. Finally, the CNN output and attribute feature vectors are combined in a fully connected layer of the size of 2 048. The final output is a full-connection layer of the size of 1 024. Dropout is applied to the fully-connected layers to reduce the risk of over-fitting.

4. Experiments

We firstly test the performance of the attribute feature classifier on the data set, and then verify the role of the attribute features in the pedestrian re-id process.

4.1 Attribute net

We assess the attribute classification method in our work on the pedestrian attributes dataset which is called PETA. PETA is an existing pedestrian data set with attribute tags. The dataset contains a total of 19 000 pedestrian images. The images data has 61 binary attribute features and 4 multi-attribute features, as shown in Fig. 10. Body part detection is used to locate the head, upper body and lower body, which can enhance the performance of pedestrian local attributes recognition. Because the PETA dataset have not any annotation of the body parts. In our paper, the Poselet [31] method is adopted for the body part detection.

To test the attribute feature extraction network, we use 11 600 images to train the network while using another 7 600 images for testing. The images are randomly chosen. In order to compare with similar works, we further use 35 attributes where the ratio of positive and negative labels exceeds 5%. As an evaluation index, we use the mean accuracy (mA) as the accuracy between the average and negative samples in the positive sample. We compare the proposed method with the three most recent methods: ACN [32], DeepMAR [33], and WPAL-GoogleNet-GMP [34]. The results are shown in Table 1. Our property network has a competitive accuracy but does not exceed the current state of the art. In Table 2, we show some of the most reliable and unreliable attributes on PETA.

![Fig. 10 Composition of PETA dataset](image-url)
4.2 Re-id based on attributes

We verify our method on two important benchmark datasets, which are used to test person re-id based video. We begin with the description of the dataset, and then introduce and analyze the experimental results.

4.2.1 Datasets

The first dataset we introduce is called iLIDS-VID [35]. It is mainly collected from the airport hall to 300 pedestrian images, as shown in Fig. 11. The length of the video from the dataset ranging from 23 – 192 frames, the average frame rate is 73 frames. This dataset is very challenging because of the changes in light and viewpoint with a complex background and occlusion. The dataset of the PRID 2011 contains 200 pairs of pedestrian images sequence taken from two adjacent cameras. The length of the images sequence from the dataset ranging from 5 – 675 frames, and the average frame rate is 100 frames, as shown in Fig. 12. Because of the relatively simple background and less occlusion, this is less challenging than the iLIDS-VID dataset. The detailed information of iLIDS-VID and PRID2011 are shown in Table 3.

![Fig. 11 Sample images in iLIDS-VID dataset](image)

![Fig. 12 Sample images in PRID 2011 dataset](image)

| Dataset   | Person | Camera | Images | Size   |
|-----------|--------|--------|--------|--------|
| iLIDS-VID | 300    | 2      | 42,495 | 128×64 |
| PRID2011  | 200    | 2      | 24,541 | 128×64 |

4.2.2 Experimental setup

We randomly sample two equal sample subsets from each pedestrian sample data set, one for training and one for testing. When testing, for each query pedestrian picture sequence, calculate the distance to each identification in the sample set to be matched and return the n identification that is in the front of the list. To be able to measure the performance of the algorithm, the rank n is used to count the experimental result data, which represents the percentage of correctly matched test sequences at the specified level. To be able to extract attribute features, 477 and 1 134 images corresponding to the relevant personnel are removed from the PRID 2011 datasets and the ILIDS-VID respectively. Next, we will randomly select 16 000 pedestrian images from the remaining pedestrian datasets for training, and use the remaining pedestrian data to collect 2 523 and 1 866 pedestrian images for testing. Our model based on the Caffe framework runs in the NVIDIA GPU 1080ti-12G. The initial learning rate is set to 0.000 1 and the momentum is 0.9. Every training needs 50 000 iterations and about 8 hours for this network model coverage. The experiments are repeated 10 times with different samples and show the performance of the average rank k marching rate.

4.2.3 Experimental results and discussion

We compare our experimental results with the most advanced algorithms as follows: a spatial-temporal appearance (STA) features and KISSME metric learning method [36], re-current CNNs (RCNNs) [37], top-push deep learning (TDL) [38], temporally aligned pooling representation (TAPR) [39], simultaneously learning intra-video and inter-video distance learning (SI2DL) [40], jointly attentive spatial-temporal pooling network (ASTPN) [41] and temporal residual learning (TRL) [42]. The final results are shown in Table 4 and Table 5. we show the recognition rates of our algorithm in the two pedestrian image datasets at rank 1, 5, 10 and 20 respectively. The cumulated matching characteristic (CMC) curves are shown in Fig.13 to report the overall performance. As can be seen from the table, the proposed method can achieve improved performances on ILIDS-VID and PRID 2011 datasets.

| Method     | r = 1 | r = 5 | r = 10 | r = 20 |
|------------|-------|-------|--------|--------|
| STA [36]   | 44.3  | 71.7  | 83.7   | 91.7   |
| RCNN [37]  | 58.3  | 84.6  | 92.0   | 96.7   |
| TDL [38]   | 56.3  | 87.6  | 95.6   | 98.3   |
| TAPR [39]  | 55.0  | 87.5  | 93.8   | 97.2   |
| SI2DL [40] | 48.7  | 81.1  | 89.2   | 97.3   |
| ASTPN [41] | 62.0  | 86.0  | 94.0   | 98.0   |
| TRL [42]   | 57.7  | 81.7  | 91.0   | 94.1   |
| Ours       | 63.2  | 86.2  | 93.7   | 98.5   |

For ILIDS-VID, our rank 1 recognition rate is 63.2%. It outperforms the second-ranked methods ASTPN by more
than 1%. The proposed model employs the structure of attribute detection, and therefore the result exceeds RCNN by 4.9%.

Table 5  Comparison of recognition rates at different ranks on PRID 2011 dataset

| Method | r = 1 | r = 5 | r = 10 | r = 20 |
|--------|-------|-------|--------|--------|
| STA [36] | 64.1 | 87.3 | 89.9 | 92.0 |
| RCNN [37] | 70.6 | 92.3 | 95.3 | 97.3 |
| TDL [38] | 56.7 | 80.0 | 87.6 | 93.6 |
| TAPR [39] | 58.2 | 85.8 | 93.4 | 97.9 |
| Si2DL [40] | 76.7 | 95.6 | 96.7 | 98.7 |
| ASTPN [41] | 74.0 | 95.0 | 99.0 | 99.0 |
| TRL [42] | 87.8 | 97.4 | 99.0 | 99.3 |
| Ours | 88.1 | 96.7 | 99.2 | 99.6 |

Fig. 13  Recognition rates of several methods on ILIDS-VID and PRID 2011 datasets

In addition, our method achieves better performance than state of the art TAPR and TRL. For PRID 2011, the performance of our result is excellent. The rank 1 achieves 88.1% using CCL and transcends the second-ranked method TRL by 0.3%. In addition, our method exceeds ASTPN by approximately 14% and Si2DL by 11.4%. It is worth noting that the introduction of global and local attributes can greatly improve the performance of re-id on both datasets. The proposed method can enhance the extracted features discriminative ability by CCL, even though it is combined with simple metric learning.

5. Conclusions

This paper proposes a deep network architecture to person re-id with attributes characteristic. The designed method shows that the attributes characteristic is an important clue to the person re-id task, and its auxiliary information can increase the description ability of pedestrian. Using the CCL training method, the proposed network combines low-level features and attributes characteristic, thus forming an end-to-end deep learning framework. Extensive experiments on two benchmark datasets demonstrate that our method is robust in attribute detection and substantially outperforms state of the art person re-id methods. Our main contributions are summarized below. (i) We propose a new method that integrates information of attributes into the network based on the CNN trained for person re-id. (ii) Using the CCL training method, the two feature training networks are combined and the combined features are mapped into the Euclidean space, thus forming an end-to-end deep learning based pedestrian recognition network framework. (iii) Experiments prove that attributes would raise the accuracy of person re-id on multiple large-scale public datasets and lead to the powerful performance of our combined model. Our future work will consider the spatial locations and correlations of attributes, which might further improve the accuracy of attribute detection.

References

[1] GONG S, CRISTANI M, YAN S, et al. Person re-identification. London: Springer, 2014.
[2] BEDAGKAR-GALA A, SHAH K S. A survey of approaches and trends in person re-identification. Image and Vision Computing, 2014, 32(4): 270 – 286.
[3] ZHAO R, OYANG W L, WANG X G. Person re-identification by saliency learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(2): 356 – 370.
[4] GHEISSARI N, SEBASTIAN T, HARTLEY R. Person re-identification using spatiotemporal appearance. Proc. of the Computer Vision and Pattern Recognition, 2006: 1528 – 1532.
[5] FARENZENA M, BAZZANI L, PERINA A, et al. Person re-identification by symmetry-driven accumulation of local features. Proc. of Computer Vision and Pattern Recognition, 2010: 2 – 8.
[6] WAN X, GE D, LI P. Person re-identification by features fusion. Proc of the IEEE Information Technology, Networking, Electronic and Automation Control Conference, 2016: 285 –
[7] KOESTINGER M, HIRZER M, WOHLHART P, et al. Large scale metric learning from equivalence constraints. Proc. of the Computer Vision and Pattern Recognition, 2012: 2288 – 2295.

[8] WEINBERGER Q K, BLITZER J, SAUL K L. Distance metric learning for large margin nearest neighbor classification. Proc. of the Neural Information Processing Systems, 2005: 1473 – 1480.

[9] ZHEN L, CHANG S Y, LIANG F, et al. Learning locally-adaptive decision functions for person verification. Proc. of the Computer Vision and Pattern Recognition, 2013: 3610 – 3617.

[10] HU J, LU J, TAN P Y. Deep transfer metric learning. IEEE Trans. on Image Processing, 2016, 25(12): 5576 – 5588.

[11] SUDOWE P, SPITZER H, LEIBE B. Person attribute recognition with a jointly-trained holistic cnn model. Proc. of the IEEE International Conference on Computer Vision Workshops, 2015: 87 – 95.

[12] HUANG J, FERIS S R, CHEN Q, et al. Cross-domain image retrieval with a dual attribute-aware ranking network. Proc. of the IEEE International Conference on Computer Vision, 2015: 1062 – 1070.

[13] MATSUKAWA T, SUZUKI E. Person re-identification using CNN features learned from combination of attributes. Proc. of the 23rd International Conference on Pattern Recognition, 2016: 2428 – 2433.

[14] LI W, ZHAO R, XIAO T, et al. Deepreid: deep filter pairing neural network for person re-identification. Proc. of the Computer Vision and Pattern Recognition, 2014: 152 – 159.

[15] CHENG D, GONG Y, ZHOU S, et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. Proc. of the Computer Vision and Pattern Recognition, 2016: 1335 – 1344.

[16] VARIOR R R, HALOI M, WANG G. Gated siamese convolutional neural network architecture for human reidentification. Proc. of the European Conference on Computer Vision, 2016: 791 – 808.

[17] FARHADI A, ENDRES I, HOIEM D, et al. Describing objects by their attributes. Proc. of the Computer Vision and Pattern Recognition, 2009: 1778 – 1785.

[18] LAMPERT H C, NICKISCH H, HARMELING S. Attribute-based classification for zero-shot visual object categorization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014, 36(3): 453 – 465.

[19] RUSSAKOVSKY O, LI F F. Attribute learning in large-scale datasets. Proc. of the European Conference on Computer Vision, 2010: 1 – 14.

[20] JAYARAMAN D, SHA F, AND GRAUMAN K. Decorrelating semantic visual attributes by resisting the urge to share. Proc. of the Computer Vision and Pattern Recognition, 2014: 1629 – 1636.

[21] ZHANG H, YANG Y, LUAN H, et al. Start from scratch: Towards automatically identifying, modeling, and naming visual attributes. Proc. of the ACM Multimedia, 2014: 187 – 196.

[22] CHEN Y, HSU H W, LIAO M H. Automatic training image acquisition and effective feature selection from community-contributed photos for facial attribute detection. IEEE Trans. on Multimedia, 2013, 15(6): 1388 – 1399.

[23] LIN C, CHEN Y, CHEN B, et al. Facial attribute space compression by latent human topic discovery. Proc. of the ACM Multimedia, 2014: 1157 – 1160.

[24] ROMERO A, GATTA C, CAMPS-VALLS G. Unsupervised deep feature extraction for remote sensing image classification.
Biographies

SUN Rui was born in 1976. He is a Ph.D. and a professor in Hefei University of Technology. His research interests are computer vision, intelligent information processing and machine learning.
E-mail: sunrui@hfut.edu.cn

HUANG Qiheng was born in 1994. He received his master degree from Hefei University of Technology. He is a R&D engineer in the China Mobile Anhui Branch. His research interests are computer vision and deep learning.
E-mail: jchqh123@163.com

FANG Wei was born in 1993. He received his master degree from Hefei University of Technology. He is a R&D engineer in the ArcSoft Corporation Limited now. His research interests are computer vision and deep learning.
E-mail: 1204764020@qq.com

ZHANG Xudong was born in 1966. He is a Ph.D. and a professor in Hefei University of Technology. His research interests are pattern recognition and intelligent information processing.
E-mail: xudong@hfut.edu.cn