Multi-label CNN Based Pedestrian Attribute Learning for Soft Biometrics

Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, Stan Z. Li*
Center for Biometrics and Security Research,
Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun Donglu, Beijing 100190, China
jianqingzhu@foxmail.com, {scliao, dyi, zlei, szli}@nlpr.ia.ac.cn

Abstract

Recently, pedestrian attributes like gender, age and clothing etc., have been used as soft biometric traits for recognizing people. Unlike existing methods that assume the independence of attributes during their prediction, we propose a multi-label convolutional neural network (MLCNN) to predict multiple attributes together in a unified framework. Firstly, a pedestrian image is roughly divided into multiple overlapping body parts, which are simultaneously integrated in the multi-label convolutional neural network. Secondly, these parts are filtered independently and aggregated in the cost layer. The cost function is a combination of multiple binary attribute classification cost functions. Moreover, we propose an attribute assisted person re-identification method, which fuses attribute distances and low-level feature distances between pairs of person images to improve person re-identification performance. Extensive experiments show: 1) the average attribute classification accuracy of the proposed method is 5.2% and 9.3% higher than the SVM-based method on three public databases, VIPeR and GRID, respectively; 2) the proposed attribute assisted person re-identification method is superior to existing approaches.

1. Introduction

Pedestrian attributes, such as *gender*, *dark hair* and *skirt* etc., have been used as soft biometric traits in the surveillance field, and has attracted a lot of attention recently. For example, pedestrian attributes can be used as useful clues for person retrieval [6, 15], subject identification [16], person recognition [5, 17], human identifying [4, 29, 31], face verification [21] and person re-identification [23]. In many real-world surveillance scenarios, cameras are usually installed at a far distance to cover wide areas, therefore pedestrians are captured with low resolution, as a result, high-quality face images are hardly attainable. However, in such scenarios pedestrian attributes still have a high application potential, because pedestrian attributes have been shown to provide several advantages beyond traditional biometrics, such as invariance to illumination and contrast [16].

There are three main challenges in pedestrian attribute classification. First, there are large intra-class variations, due to diverse clothing appearances, various illumination conditions and different camera views. As shown in Figure 1, the *backpack* annotated samples in the VIPeR [10] database captured with different cameras have drastic appearance variations. Second, pedestrian attributes have complex localizing characteristics, which means that some attributes can only be recognized in some certain or uncertain local body areas. For example, *long hair* is most relevant with the head and shoulders areas; *satchel* (see Figure 1) may appear in either left of right side of the image, with uncertain height. As a result, pedestrian attribute feature extraction is very difficult. Third, pedestrian attribute classification is a multi-label classification problem instead of a multi-class classification problem, because pedestrian attributes are not completely mutually exclusive. Therefore, most of the existing multi-class classification algorithms are not applicable, and multi-label classification has its own challenge.

The most popular approach for attribute prediction is the one using hand-crafted features and SVM based individual attribute classifier [1, 8, 16, 22, 23], which cannot solve the above mentioned challenges successfully because hand-crafted features have limited representation ability for large intra-class variations, and independent SVM classifiers cannot investigate interactions between attributes. In this paper, we present a comprehensive study on pedestrian attribute classification. We solve the multi-attribute classification problem with a multi-label convolutional neural network (MLCNN). The multi-label convolutional neural network is trained from raw pixels rather than hand-crafted features and is able to simultaneously recognize multiple attributes, which achieves higher accuracy than the SVM-based attribute classifiers proposed in [1, 22, 23]. Moreover,

*Corresponding author.

Figure 1. Annotated sample images from the VIPeR and GRID databases.

due to the improvement of the proposed attribute classification method, fusing attribute distances and low-level feature distances between pairs of person images leads to a better person re-identification method than existing approaches.

2. Related Work

The study of attributes is receiving more and more interests, since attributes are helpful to infer high-level semantic knowledge. There are many computer vision applications based on attributes, such as face verification [21], image retrieval [34], clothing description [3], human attribute recognition [35]. The most related work are [1, 22, 23, 26], where the attributes are automatically predicted from a low resolution pedestrian image. In these works, Layne et al. [23] defined 15 human-understandable pedestrian attributes such as *male*, *longhair*, *backpacks*, *headphones* and *clothing* on the VIPeR [10] and i-LIDS databases. They further provided 21 attribute annotations on the VIPeR, PRID [13] and GRID [28] databases in [22]. Figure 1 shows some examples of annotated images. For attribute classification, all of these works train SVM classifier for each attribute independently, which ignores the interaction between different attributes.

Convolutional neural networks (CNN) [7, 12, 20, 24, 25] have been used in many image-related applications and exhibited good performances. However, most of these works are concerned with single-label image classification, and each image in the dataset only contains one prominent object class. Recently, Gong et. al [9] proposed a multi-label deep convolutional ranking net to address the multi-label annotation problem. They adopted the architecture proposed in [20] as basic framework and redesigned a multi-label ranking cost layer for multi-label prediction tasks.

There are several person re-identification methods using attribute information. In [1], gallery images are ranked corresponding low-level feature distances first, then locally adjusted by attribute distances. [23] fuses attribute information on score level, which calculates an attribute distance and fuse it with a low-level feature distance to form the fusion distance. [22] utilizes attribute information on feature level, which concatenates the predicted attribute scores and low-level features and learns discriminative metric on the concatenated features. The latent SVM based person re-identification approach [26] embeds clothing attributes as latent real-value variables via a continuous latent SVM

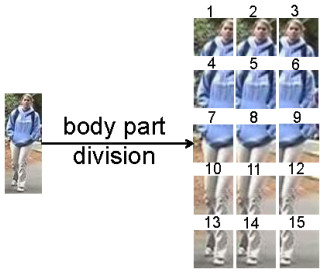

Figure 2. One person is divided into 15 overlapping body parts.

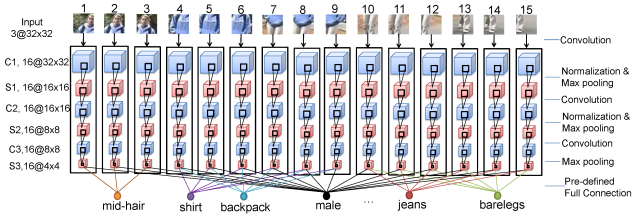

Figure 3. The structure of the multi-label convolutional neural network (MLCNN) used in our method.

[33] framework. It implicitly describes the relations among the low-level part features, mid-level clothing attributes and high-level re-identification labels of person pairs.

3. Pedestrian Attribute Classifier Training

3.1. Body Part Division

Because of body movements, commonly used holistic feature representation methods suffer from the pose misalignments. Besides that, some attributes have local characteristic. For example, *long hair* is most relevant with head and shoulders area; *backpack* is most likely to appear in upper torso region; *jeans* appears in lower part of body. Considering these factors, in [3, 26], part detection is first applied to locate body parts and low-level features are extracted from the detected regions. However, part detection itself, is a challenging problem, due to the geometric variation such as articulation and viewpoint changes as well as the appearance variation of the parts arisen from versatile clothing types. We do not use a body part detector, but roughly divide a pedestrian image into 15 overlapping 32×32 sized parts, as shown in Figure 2. The steps of horizontal and vertical axises are 8 pixels and 24 pixels, respectively.

3.2. Multi-label Convolutional Neural Network

After body part division, multiple parts are integrated to a multi-label convolutional neural network (MLCNN) at the same time, as shown in Figure 3. Each part is filtered independently. All convolutional and pooling layers are designed with 16 channels. The cross channel normalization unit [20] is used in the S1 and S2 pooling layers. The filter sizes of C1, C2 and C3 layer are 7×7, 5×5 and 3×3, respectively. The ReLU neuron [20] is used as activation function for each layer. Dropout [12] layers follow each of the fully connected layers with a dropout ratio of 0.5.

Table 1. The pre-defined connection relationships between parts and attributes.

attribute	parts	attribute	parts	attribute	parts
male	1-15	redshirt	4-9	skirt	10-15
midhair	1-3	blueshirt	4-9	barelegs	10-15
darkhair	1-3	nocoats	4-9	shorts	10-15
bald	1-3	patterned	4-9	lightbottoms	10-15
darkshirt	4-9	hassatchel	4-9	darkbottoms	10-15
lightshirt	4-9	hasbackpack	4-9	jeans	10-15
greenshirt	4-9	hashandbag carrierbag	7-12	notlightdark jeanscolour	10-15

Moreover, considering some attributes (e.g. *bald*, *longhair*, *jeans*) are unlikely to appear outside of their expected parts, we pre-define connections between attributes and parts. As shown in Figure 3, *male* is fully connected with all parts, *mid-hair* is fully connected with part 1-part 3, *shirt* is fully connected with part 4-part 9 and *jeans* is fully connected with part 9-part 15, etc. Table 1 lists the pre-defined connection relationships between parts and attributes.

3.3. Cost Function and Learning

Since attributes are not completely mutually exclusive, multi-attribute prediction is a multi-label classification problem essentially. The last layer of the proposed MLCNN structure is different from the CNN used for single-label classification which usually only includes one cost function. In order to make our MLCNN to predict all attribute classifiers together, we sum all attribute classification cost together. Following [9, 11], we use the softmax function [2] for each attribute prediction. The cost function of our multi-label convolutional neural network (MLCNN) is defined as follows:

$$F = \sum_{k=1}^{K} \lambda_k G_k. \tag{1}$$

where G_k is the loss of the k-th attribute; K is the total number of the attributes; $\lambda_k \geq 0$ is a parameter used to control the contribution of the k-th attribute. In our experiments, we set $\lambda_k = \frac{1}{K}$ and define G_k as follows:

$$G_k = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M^k} 1\{y_n^k = m\} \cdot \log \frac{e^{(w_m^k)^T \cdot x_n^k}}{\sum_{m=1}^{M^k} e^{(w_m^k)^T \cdot x_n^k}}, \tag{2}$$

where $\{x_n^k, y_n^k\}$ represents a training sample and y_n^k is k-th attribute label of n-th sample x_n^k; N represents the number of training sample and M^k represents the class number of k-th attribute; $1\{\cdot\}$ is a indicator function. To avoid bias due to imbalanced data, we further extend Eq. (2) as follows:

$$G_k = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M^k} 1\{y_n^k = m\} \cdot \beta_m^k \cdot \log \frac{e^{(w_m^k)^T \cdot x_n^k}}{\sum_{m=1}^{M^k} e^{(w_m^k)^T \cdot x_n^k}},$$
$$\beta_m^k = \frac{\frac{1}{N_m^k}}{\sum_{l=1}^{M^k} \frac{1}{N_l^k}}, \tag{3}$$

where N_m^k is number of samples holding m-th class label of k-th attriube and it meets $\sum_{m=1}^{M^k} N_m^k = N^k$. Back propagation (BP) [24] is used to learn the parameters of the MLCNN and there are many public CNN tools, such as cuda-convnet [20] and Caffe [18].

4. Attribute Assisted Person Re-identification

Person re-identification is to recognize individuals through person images captured from multiple non-overlapping cameras. In practice, person re-identification is aimed to return a ranked list. In the ranked list, gallery images are ranked according to distances between gallery images and a probe image. The higher true gallery images are ranked, the better performance will be achieved.

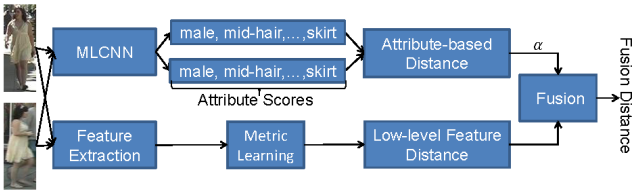

Figure 4. The framework of attribute assisted person re-identification.

As shown in Figure 4, a fusion distance consists of attribute distance and low-level feature distance. Assume that x_p is a probe image and x_g is a gallery image, then the fusion distance $S_f(x_p, x_g)$ is defined as follows:

$$S_f(x_p, x_g) = S_l(x_p, x_g) + \alpha S_a(x_p, x_g), \tag{4}$$

$$S_l(x_p, x_g) = (L_p - L_g)^T M (L_p - L_g), \tag{5}$$

$$S_a(x_p, x_g) = (A_p - A_g)^T (A_p - A_g), \tag{6}$$

where $S_l(x_p, x_g)$ is the low-level feature distance between x_p and x_g; $S_a(x_p, x_g)$ is the attribute distance between x_p and x_g; L_p and L_g are the low-features of x_p and x_g; M is a metric learned by KISSME [19]; A_p and A_g are the attribute scores predicted by MLCNN. α is set as 0.5 in our experiments. Once the fused distances between gallery images and a probe image are obtained, the rest work of person re-identification is to rank gallery images according to the fused distances.

5. Experiments

Two challenging databases are used to validate our algorithm, VIPeR [10] and GRID [28]. VIPeR [10] contains 632 pedestrian image pairs from two cameras with different viewpoint, pose and lighting. GRID [28] provides 1,275 pedestrian images captured by cameras installed in a busy underground station and only 250 person image pairs of them captured by two cameras with different viewpoints. GRID is challenging due to variations of pose, colours, lighting changes, as well as poor image quality caused by low spatial resolution. The two databases are annotated with 21 attributes and the annotations are provided by [22]. Figure 1 shows some annotated samples.

We conducted both attribute classification and person re-identification experiments. For attribute classification, we compared the proposed MLCNN method to SVM and CNN. Besides, for MLCNN we evaluated both **all** body based method (denoted as mlcnn-a) and **part** based method (denoted as mlcnn-p).

5.1. Setup

All images are scaled into 128×48 pixels. For each database, we split it into non-overlapping training and testing sets randomly and the sizes of training and testing are set the same. We repeat the process 11 times and got 11 splits. The 11-th split is used for parameter tuning, such as the number of epoch, learning rate, weight decay and so on. The other 10 splits are used for reporting the results. The features used for training SVM-based attribute classifiers are the same with [22], which have 2784 dimensions. For person re-identification, we used a 12,750-dimentional descriptor proposed in [27], which fuses HSV and MB-LBP region histograms. Before KISSME [19] metric learning, we project the feature descriptors into a 100 and 40 dimensional subspaces by PCA for the VIPeR and GRID databases, respectively.

5.2. Attribute Classification

We retrain the SVM-based attribute classifiers proposed in [22] and compare it with our method. Following the evaluation protocol of [22], we report the average recall rate of each attribute classifier with default thresholds (0 for SVM, 0.5 for MLCNN). From Table 2 and Table 3, we can see that the average accuracy rates of mlcnn-a and mlcnn-p are higher than that of the SVM-based method [22] for the two databases. The average accuracy rates of the mlcnn-p classifiers are 5.2% and 9.3% higher than that of the SVM based classifiers on VIPeR and GRID, respectively.

In order to evaluate the performance of attribute classification more comprehensively, we further report the average recall rate when the false positive rate (FPR) is at 20% (shown in Table 2 and Table 3). One can see that the average recall rates of mlcnn-a and mlcnn-p are higher than that of the SVM-based method [22]. The mlcnn-p achieves the best performance. Compared with the SVM-based method, the mlcnn-p obtains 9.4% and 8.2% improvements on average recall rates for VIPeR and GRID databases, respectively. Furthermore, the overall ROC curves are shown on Figure 5 for the two databases. Based on Figure 5, we can clearly see that the mlcnn-a/mlcnn-p has better performance than the SVM-based classifier. These results demonstrate our MLCNN is superior to the SVM-based method. Moreover, we can find that mlcnn-p is better than mlcnn-a in most cases. Especially, for those local attributes *midhair* and *darkhair*, mlcnn-p is much better than mlcnn-a. This result demonstrates the pre-defined relationships between attributes and body parts are effective for attribute classification on the two databases.

Additionally, we independently train each attribute's classifier by using CNN [20] on the VIPER database and compare the performance with mlcnn. The structure of each independent attribute CNN is the same with mlcnn and the pre-defined part-attribute relationship (Table 1) is also used.

Table 2. Comparison of attribute classification performance on VIPeR.

attribute	average±std Accuracy Rate (%)			average±std Recall Rate (%) @ FPR=20%		
	SVM	mlcnn-a	mlcnn-p	SVM	mlcnn-a	mlcnn-p
male	66.5±1.1	68.4 ± 1.3	**69.6** ± 2.6	48.2±3.5	55.1±2.6	**57.2**±3.7
midhair	64.1±2.3	72.2 ± 2.1	**76.1** ± 1.8	43.0±3.9	52.2±3.0	**63.5**±4.2
darkhair	63.9±1.8	69.8 ± 1.5	**73.1** ± 2.1	39.6±2.7	50.9±3.1	**58.4**±5.8
darkshirt	**84.2**±0.9	83.4 ± 1.1	82.3± 1.4	**87.5**±1.2	86.3±2.0	85.8±2.1
lightshirt	**83.7** ±1.0	83.4 ± 1.4	83.0± 1.2	**87.8**±1.3	86.5±2.9	85.3±2.3
greenshirt	71.4±5.2	66.4 ± 5.2	**75.9**± 5.9	54.3± 9.5	51.5±12.2	**69.4**±8.0
redshirt	85.5±2.3	90.7 ± 1.6	**91.9**± 1.0	88.4± 3.9	86.0±5.7	**88.9**±4.8
blueshirt	**73.0**±5.2	70.2 ± 5.3	69.1± 3.3	60.8± 3.9	66.5±3.2	**70.8**±5.1
nocoats	70.6±1.9	69.9 ± 1.6	**71.3**± 0.8	**59.3**± 2.4	55.9±3.2	57.2±3.2
patterned	46.9±15.1	56.0 ± 4.9	**57.9**± 9.2	26.3± 6.0	28.5±9.7	**41.0**±9.0
lightbottoms	74.7±1.2	76.2 ± 1.6	**76.4**± 1.2	69.5± 3.0	73.0±3.6	**73.3**±2.5
darkbottoms	75.7±1.7	77.4 ± 1.5	**78.4**± 0.7	70.2± 4.7	74.5±3.7	**76.2**±1.9
notlightdark jeanscolour	70.3±7.3	**90.8** ± 1.9	90.7 ± 2.0	57.2± 7.9	73.4±6.8	**78.6**±7.5
jeans	76.4±1.3	76.0 ± 1.5	**77.5**± 0.6	72.7± 3.4	72.0±3.2	**74.7**±2.8
skirt	63.6±8.8	**78.2** ± 3.5	78.1± 3.5	40.7± 13.9	50.0±10.1	**60.7**±9.9
barelegs	75.6±3.8	81.7 ± 1.4	**84.1**± 1.1	68.7± 6.5	73.3±5.9	**85.4**±4.5
shorts	70.4±5.2	77.3 ± 2.7	**81.7** ± 1.3	59.8± 6.5	62.6±7.4	**82.9**±4.7
hassatchel	47.8±4.8	55.7 ± 3.3	**57.8** ± 2.7	22.0± 4.9	29.4±2.9	**31.7**±4.3
hashandbag carrierbag	**45.3**±3.8	45.1 ± 6.2	42.0 ± 6.5	17.4± 3.5	**22.0**±6.5	18.5±5.8
hasbackpack	**67.5**±1.4	63.8 ± 2.7	64.9 ± 1.2	47.9± 4.7	43.4±3.6	**49.9**±3.7
average±std	68.9±1.1	72.6± 0.6	**74.1** ± 1.0	56.1± 1.3	59.7±1.4	**65.5**±1.5

Table 3. Comparison of attribute classification performance on GRID.

attribute	average±std Accuracy Rate (%)			average±std Recall Rate (%) @ FPR=20%		
	SVM	mlcnn-a	mlcnn-p	SVM	mlcnn-a	mlcnn-p
male	63.2± 2.9	65.8±2.7	**68.4**±1.8	42.8±8.2	50.2±6.7	**52.8**±4.9
midhair	61.1± 2.8	68.0±2.4	**72.4**±3.4	38.4±8.5	45.3±6.3	**60.9**±7.8
darkhair	59.6 ± 5.0	66.0±2.9	**71.8**±3.6	37.6±9.3	41.7±6.1	**58.3**±6.2
darkshirt	77.5± 2.1	79.6±2.1	**81.2**±1.9	78.1±4.5	80.5±5.6	**84.4**±5.9
redshirt	74.3± 4.9	89.2±2.7	**90.4**±2.9	65.8±10.4	82.0±8.6	**87.3**±7.1
blueshirt	77.8± 5.9	**85.8**±2.2	84.8± 2.8	70.8±8.9	76.4±10.2	**85.2**±6.9
patterned	58.5± 13.7	72.4 ±4.0	**74.3**±3.3	38.3±13.7	32.3±10.0	**44.7**±13.3
lightbottoms	**83.6**± 2.3	81.3 ±1.3	83.5 ± 2.9	87.0±4.2	83.7±3.3	**86.8**±5.2
darkbottoms	**83.8**± 2.4	81.4 ±2.9	**83.8**± 2.6	86.8±3.7	83.6±3.9	**86.6**±4.9
jeans	60.6± 3.1	59.6 ±3.1	**62.4**± 1.8	40.7±5.4	36.2±5.1	**42.2**±6.9
skirt	27.0± 31.7	78.0 ±2.9	**73.8**± 4.9	17.3±5.7	43.3±9.2	**44.4**±13.6
barelegs	62.0 ± 5.5	**76.4**±3.7	**76.4**± 2.4	40.0±7.6	57.4±12.7	**65.4**±5.7
shorts	62.3 ± 5.5	**75.5**±3.2	67.4± 5.0	39.5±10.5	**44.3**±6.9	22.0±5.5
hassatchel	55.4± 1.8	**57.2**±2.6	55.8± 3.6	**29.6**±3.6	28.0±4.6	26.9±4.8
hashandbag carrierbag	54.6± 8.8	**62.5** ±3.4	61.8± 2.8	30.1±5.1	34.1±6.5	**34.7**±8.5
hasbackpack	61.8 ± 2.4	**66.3**±2.5	63.1± 3.4	**43.3**±3.4	31.7±7.5	33.7±6.2
average±std	63.9± 2.3	72.8 ±0.9	**73.2** ± 0.7	49.1±1.8	53.2±1.7	**57.3**±0.9

We find that the average performance of independent attribute CNNs is nearly the same with mlcnn-p. However, the training and test of mlcnn-p are much faster than independent attribute CNNs. For each trial on the VIPER database, the training time of mlcnn-p and independent attribute CNNs are 28.1 mins and 146.4 mins and the test time of mlcnn-p and independent attribute CNNs are 3.6 mins and 22.5 mins, respectively.

5.3. Person Re-identification

We further investigate whether our MLCNN predicted attribute scores can further used as soft-biometric cues to improve person re-identification performance. The cumulative matching characteristic (CMC) curve [19, 22, 37] is used to measure the performance of person re-identification.

Firstly, we rank gallery images only according to the Euclidean attribute distances calculated by Eq. (6). The results

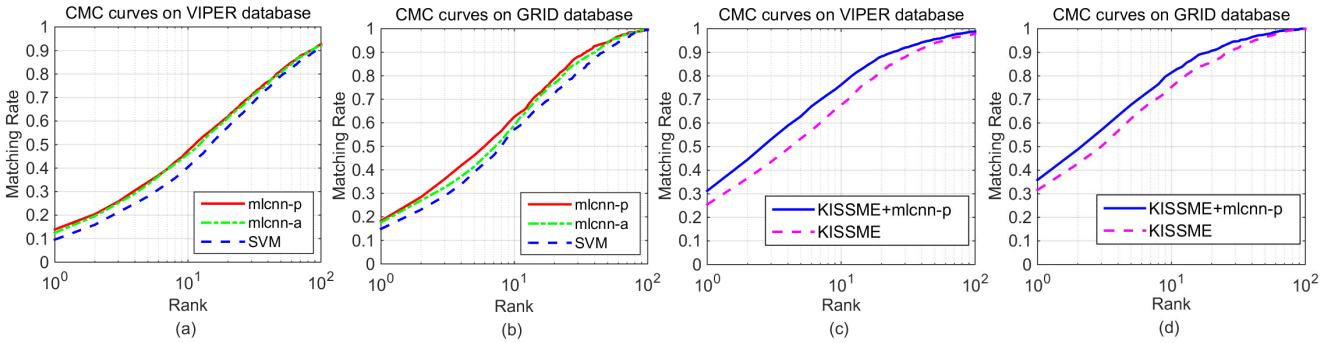

Figure 6. (a) and (b) are CMC curves on VIPeR and GRID by using only attribute distances (Eq. (6)). (c) and (d) are CMC curves on VIPeR and GRID by using the low-level distances (KISSME, Eq. (5)) and the fusion distances (KISSME+mlcnn-p, Eq. (4)).

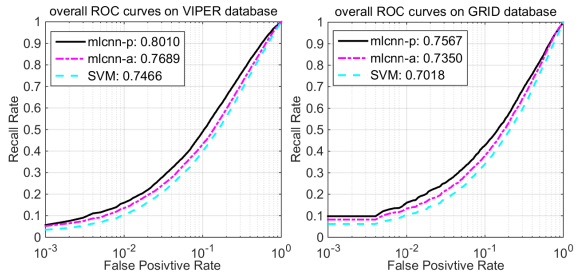

Figure 5. Overall ROC curve comparison on VIPeR and GRID. The float number in the legend rectangle represents the area under the curve (AUC).

are shown in Figure 6(a) and Figure 6(b) and Table 4. We can find that the person re-identification performances of using mlcnn-a/mlcnn-p predicted attribute scores are consistently better than that of using the SVM-based classifier predicted attribute scores. This shows that the attribute classification improvements of our MLCNN method are useful for boosting person re-identification performance.

Secondly, we summarize the results of the attribute assisted person re-identification method in Figure 6(c), Figure 6(d) and Table 5. We can find that introducing mlcnn-p prediction attribute scores is able to improve the person re-identification performance on the two databases. As shown in Table 5, compared with KISSME without fusion attribute information, our method achieves 5.79% and 4.24% improvements of rank-1 recognition rate on the VIPeR and GRID databases, respectively.

Finally, we compare our attribute assisted person re-identification method with state-of-the-art algorithms on the VIPeR database, because only the VIPeR database is completely annotated with attributes and able to meet the same evaluation protocol with state-of-art methods [14, 19, 30, 32, 36, 37]. From Figure 7 and Table 6, we can see that the proposed method outperforms most of the compared methods. Specially, before rank 25, our method outperforms the state-of-the-art method [32], though [32] has better results in higher ranks. Moreover, From Table 6, we can find that the rank-1 recognition rate of our method achieves 31.23%, while the recognition rates at rank 1 of the compared state-of-the-art methods are consistently lower than 30%.

Table 4. Comparison of attribute score based person re-identification performance on the VIPeR and GRID databases.

database	rank-1			rank-5			rank-10			rank-25		
	SVM	mlcnn-a	mlcnn-p	SVM	mlcnn-a	mlcnn-p	SVM	mlcnn-a	mlcnn-p	SVM	mlcnn-a	mlcnn-p
VIPeR	9.59	12.47	**13.89**	27.94	33.07	**34.02**	40.41	45.98	**47.41**	62.72	66.04	**67.25**
GRID	14.88	17.68	**18.32**	39.04	41.36	**46.16**	57.20	58.88	**62.56**	77.68	82.56	**84.16**

Table 5. The comparison of recognition rate (%) the KISSME without and with the fusion of the mlcnn-p classifier on the VIPeR and GRID databases.

database	rank-1		rank-5		rank-10		rank-25	
	KISSME	KISSME+ mlcnn-p	KISSME	KISSME+ mlcnn-p	KISSME	KISSME+ mlcnn-p	KISSME	KISSME+ mlcnn-p
VIPeR	25.44	**31.23**	53.42	**62.85**	67.66	**76.23**	85.54	**90.28**
GRID	31.60	**35.84**	61.84	**67.76**	75.28	**81.36**	89.28	**92.96**

Table 6. The comparison of recognition rate (%) between our method and state-of-the-art methods.

method	rank-1	rank-5	rank-10	rank-15	rank-20	rank-25	rank-30	rank-50
RDC [37]	15.66	38.42	53.86	N/A	70.09	N/A	N/A	N/A
DML [32]	28.23	59.27	73.45	81.20	86.39	89.53	**92.28**	**96.68**
LFDA [30]	24.18	N/A	67.12	N/A	N/A	85.10	N/A	94.12
RPLM [14]	27	N/A	69	N/A	83	N/A	N/A	95
Salience [36]	26.74	50.70	62.37	N/A	76.36	N/A	N/A	N/A
Ours	**31.23**	**62.85**	**76.23**	**83.73**	**88.26**	**90.28**	91.99	95.54

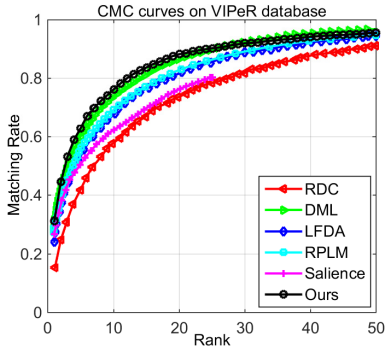

Figure 7. The comparison of our method and state-of-the-arts on VIPeR.

6. Conclusion

In this paper, a multi-label convolutional neural network (MLCNN) for pedestrian attribute classification and soft biometrics have been proposed. Moreover, we have proposed an attribute assisted person re-identification method which fuses attribute distances and low-feature based distances between pairs of person images to improve the performance of person re-identification. Experimental results on two public databases, VIPeR and GRID have well

demonstrated the effectiveness of both the MLCNN attribute classification method and the attributes assisted person re-identification method.

7. Acknowledgement

This work was supported by the Chinese National Natural Science Foundation Projects #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

References

[1] L. An, X. Chen, M. Kafai, S. Yang, and B. Bhanu. Improving person re-identification by soft biometrics based reranking. In *Distributed Smart Cameras (ICDSC), 2013 Seventh International Conference on*, pages 1–6. IEEE, 2013.

[2] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[3] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European Conference on Computer Vision*, pages 609–623. 2012.

[4] A. Dantcheva, J. Dugelay, and P. Elia. Soft biometrics systems: Reliability and asymptotic bounds. In *Fourth IEEE Int. Conference on Biometrics: Theory Applications and Systems, BTAS*, 2010.

[5] A. Dantcheva, J.-L. Dugelay, and P. Elia. Person recognition using a bag of facial soft biometrics (bofsb). In *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, pages 511–516. IEEE, 2010.

[6] A. Dantcheva, A. Singh, P. Elia, and J. Dugelay. Search pruning in video surveillance systems: Efficiency-reliability tradeoff. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1356–1363. IEEE, 2011.

[7] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pages 1223–1231, 2012.

[8] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. *Proceedings of ACM Multimedia (ACM MM)*, 2014.

[9] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

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

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316. IEEE, 2009.

[12] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[13] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. Springer, 2011.

[14] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Computer Vision–ECCV 2012*, pages 780–793. Springer, 2012.

[15] E. S. Jaha and M. S. Nixon. Analysing soft clothing biometrics for retrieval. 2014.

[16] E. S. Jaha and M. S. Nixon. Soft biometrics for subject identification using clothing attributes. 2014.

[17] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Biometric Authentication*, pages 731–738. Springer, 2004.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

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

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

[21] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE Conference on International Conference on Computer Vision*, pages 365–372, 2009.

[22] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014.

[23] R. Layne, T. M. Hospedales, S. Gong, et al. Person re-identification by attributes. In *BMVC*, volume 2, page 3, 2012.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[25] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[26] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan. Clothing attributes assisted person re-identification.

[27] S. Liao, Z. Mo, Y. Hu, and S. Z. Li. Open-set person re-identification. *arXiv preprint arXiv:1408.0872*, 2014.

[28] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 391–401. Springer, 2012.

[29] E. Martinson, W. Lawson, and J. G. Trafton. Identifying people with soft-biometrics at fleet week. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pages 49–56. IEEE, 2013.

[30] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3318–3325. IEEE, 2013.

[31] D. Reid, M. Nixon, and S. Stevenage. Soft biometrics; human identification using comparative descriptions. 2013.

[32] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *In Proceedings of International Conference on Pattern Recognition*. IEEE, 2014.

[33] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176. ACM, 2009.

[34] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2949–2956. IEEE, 2012.

[35] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. *arXiv preprint arXiv:1311.5591*, 2013.

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

[37] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):653–668, 2013.