doi:10.3772/j.issn.1006-6748.2018.01.007

Pedestrian attribute classification with multi-scale and multi-label convolutional neural networks①

Zhu Jianqing(朱建清)*, Zeng Huanqiang**, Zhang Yuzhao②*, Zheng Lixin*, Cai Canhui*
(*Fujian Academic Engineering Research Centre in Industrial Intellectual Techniques and Systems, College of Engineering, Huaqiao University, Quanzhou 362021, P. R. China)
(**School of Information Science and Engineering, Huaqiao University, Xiamen 361021, P. R. China)

Abstract

Pedestrian attribute classification from a pedestrian image captured in surveillance scenarios is challenging due to diverse clothing appearances, varied poses and different camera views. A multi-scale and multi-label convolutional neural network (MSMLCNN) is proposed to predict multiple pedestrian attributes simultaneously. The pedestrian attribute classification problem is firstly transformed into a multi-label problem including multiple binary attributes needed to be classified. Then, the multi-label problem is solved by fully connecting all binary attributes to multi-scale features with logistic regression functions. Moreover, the multi-scale features are obtained by concatenating those featured maps produced from multiple pooling layers of the MSMLCNN at different scales. Extensive experiment results show that the proposed MSMLCNN outperforms state-of-the-art pedestrian attribute classification methods with a large margin.

Key words: pedestrian attribute classification, multi-scale features, multi-label classification, convolutional neural network (CNN)

0 Introduction

At present, research of pedestrian attribute classification has attracted a lot of attention. Pedestrian attributes, such as *gender*, *dark hair* and *skirt*, can be used as soft biometric traits in the surveillance field for public security. For example, pedestrian attributes are useful clues for person retrieval[1,2], subject identification[3], human identification[4,5] and person re-identification[6,7]. In practical surveillance scenarios, pedestrian attribute classification is a challenging task in computer vision, since pedestrian images are usually of low resolution, blurred and partially occluded and contain variations of illumination and viewpoint. Therefore, how to develop an effective pedestrian attribute classification method becomes a very challenging and desirable topic.

The most popular method for pedestrian attribute classification is the one using hand-crafted features (e.g., MBLBP[8], RGB, HSV and YCbCr color histograms, Gabor and Schmid features[6]) and support vector machine (SVM) based attribute independent classifiers[3,6,9-11], which cannot fully solve the pedestrian attribute classification problem. Because hand-crafted features have a limited representation ability for large appearance variations, and attribute independent SVM classifiers cannot investigate interactions of different attributes. Moreover, along with the increasing numbers of pedestrian attributes, training SVM-based attribute classifier one by one is very tedious.

In this study, a multi-scale and multi-label convolutional neural network (MSMLCNN) is proposed to solve the pedestrian attribute classification problem. Following the VGGNet architecture[12] that applies small sized filters for each convolutional layer and emplaces multiple convolutional layers before one pooling layer, a very deep network with a strong feature learning ability can be obtained in this paper. However, it is difficult to train this very deep network, because the gradient vanishing problem[13] may appear in the back propagation process. Moreover, for those attributes

① Supported by the National Natural Science Foundation of China (No. 61602191, 61672521, 61375037, 61473291, 61572501, 61572536, 61502491, 61372107, 61401167), the Natural Science Foundation of Fujian Province (No. 2016J01308), the Scientific and Technology Funds of Quanzhou (No. 2015Z114), the Scientific and Technology Funds of Xiamen (No. 3502Z20173045), the Promotion Program for Young and Middle aged Teacher in Science and Technology Research of Huaqiao University (No. ZQN-PY418, ZQN-YX403) and the Scientific Research Funds of Huaqiao University (No. 16BS108)

② To whom correspondence should be addressed. E-mail: zyz@hqu.edu.cn

Received on Apr. 6, 2017

with complex localizing characteristics and different scales, the way of only using the features learned in the last layer is not completely suitable. Because features learned in the last layer are too global to some local attributes, such as *has sunglasses*, *upper body vneck* and *footwear sandals*. Therefore, the proposed MSML-CNN is designed by fully connecting each attribute with multiple pooling layers at different scales, which not only adds the supervisory signal to multiple intermediate layers, but also combines local and global features for the attribute classification.

The rest of this paper is organized as follows. Section 1 summarizes the related work. Section 2 introduces the proposed multi-scale and multi-label convolutional neural network. Section 3 presents experimental results to validate the superiority of the proposed method. Section 4 concludes this work.

1 Related work

1.1 Attribute pedestrian database

Recently, several public attribute pedestrian databases have been released, such as VIPeR[14], PRID[15], GRID[16], APiS[8] and PETA[11]. In terms of the number of annotated pedestrian attributes, VIPeR is firstly annotated with 15 attributes by Layne, et al.[6]. They annotated VIPeR, PRID and GRID with 21 attributes in their further work[9]. APiS annotated with 15 attributes by Zhu, et al.[8]. PETA is a large database, including 65 attribute annotations. Fig. 1 shows some annotated sample images selected from the PETA[11] database. In terms of the number of images, VIPeR, PRID and GRID are small databases and each one contains less than 1500 images. APiS includes 3661 images and PETA consists of 19000 images. It can be found that more and more databases have been released, and both the number of attributes and the number of images are increasing. This illustrates the research of pedestrian attribute classification is attracting more and more interest and attention.

Fig. 1 Annotated sample images selected from the PETA[11] database.

1.2 Pedestrian attribute classification

The most popular approach for pedestrian attribute classification is to train each attribute classifier independently on hand-crafted features. In terms of hand-crafted features, there are many local features possible be applied to describe pedestrian images for different attribute classifications, such as MBLBP[8], RGB, HSV and YCbCr color histograms, Gabor and Schmid[6] features. Moreover, sparse feature representation methods[17,18] also can be used to represent pedestrian images.

In terms of attribute classifiers, support vector machines (SVMs) are most commonly used. For example, support vector machines (SVMs) were applied to train attribute independent classifiers in Refs[3, 6, 9, 10]. Moreover, the gentle AdaBoost[19] algorithm was utilized to train each attribute's classifier independently in Ref. [8]. If the number of pedestrian attributes is small, these straight forward methods are able to train pedestrian attribute independent classifiers conveniently. However, when the number of pedestrian attributes is huge, the one by one attribute independent training progress is too tedious for human. In addition, these methods still leave a room for improving the accuracy of pedestrian attribute classification, because they do not take the interactions of different attributes into account.

Considering that pedestrian attribute classification is a multi-label classification problem, rather than a multi-classification problem[20], there are some methods learning interaction models of different attributes to improve the performance of pedestrian attribute classification. For example, Chen, et al.[21] applied a conditional random field (CRF) to learn an attribute interaction model. Deng, et al.[11] built an undirected graph with a Markov random field (MRF) to model the relationships of different attributes. In the previous work[22], pedestrian attribute classification was improved by weighting interactions from other attributes.

1.3 Convolutional neural network

Convolutional neural networks (CNNs)[23,24] have been used in many image-related applications and exhibited good performances. Krizhevsky, et al.[25] applied AlexNet for image classification and it outperformed many state-of-the-art methods on the ImageNet database. Donahue, et al.[26] and Razavian, et al.[27] demonstrated that off-the-shelf features learned by a CNN pre-trained on the ImageNet database could be effectively adopted to attribute classifications. Sun, et al.[28] proposed a CNN named DeepID to learn a set of high-level feature representations for face verification. DeepID achieves a 97.45% face verification accuracy on the LFW database and it is almost as good as the human performance of 97.53%. Based on DeepID, DeepID2[29] and DeepID3[30], further improvement is got for the face verification accuracy on the LFW database. Gong, et al.[31] proposed a multi-label deep convolutional ranking network to address the multi-label image annotation problem. They adopted the architecture proposed in Ref. [25] as a basic architecture and redesigned a multi-label ranking cost layer for multi-label prediction tasks. Zhu, et al.[32] proposed a multi-label convolutional neural network for pedestrian attribute classification, which learns single scale features for all attributes.

The most popular components in the recent research of CNN include rectified linear unit (ReLU) neuron[33], dropout[34], batch normalization[35], adding supervisory signals to intermediate layers[36], multi-scale fully connection[37], joint identification-verification cost function[29] and small filter and very deep architecture[12], ResNet[38,39] and DenseNet[40].

2 Multi-scale and multi-label convolutional neural network

2.1 Network architecture

As shown in Fig. 2, the proposed multi-scale and multi-label convolutional neural network (MSMLCNN) uses a VGGNet[12] architecture which emplaces two continuous convolutional layers before the first two max pooling layers. Moreover, different from VGGNet[12], the proposed MSMLCNN adds supervisory signals to the last three pooling layers (i.e., layers 8, 10, 12). The VGGNet architecture has shown that continuous convolutions are able to learn features with larger receptive fields and obtain more complex nonlinearity while restrict the number of parameters.

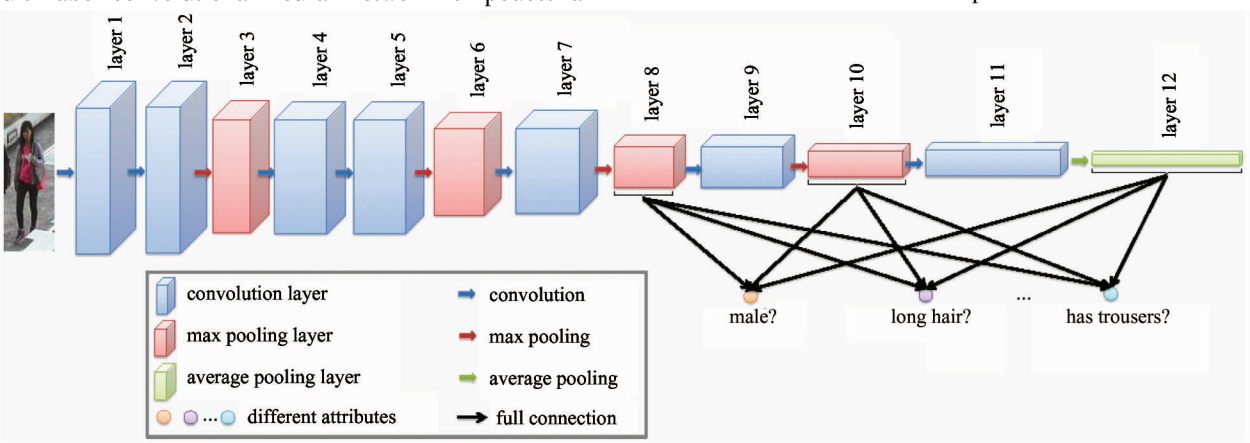

Fig. 2 The architecture of the deep multi-scale and multi-label convolutional neural network (MSMLCNN) used in the proposed method

Refs[30,36] have shown the benefits of introducing supervisory signals to multiple layers in two aspects. First, it is useful to learn more discriminative mid-level features. Second, it is able to avoid gradient vanish to make the optimization of a very deep neural network easier. However, in Refs[30,36], when training a deep CNN, supervisory signals were added by connecting some intermediate layers with multiple cost functions configured with different weights. In the testing process, only the prediction model learned on the last layer is used for predicting, while those prediction models learned on the intermediate layers are discarded. This way of adding supervisory signals is not completely suitable for pedestrian attribute classification. The reasons are listed as follows.

First, it is very difficult to assign suitable weights to different attribute cost functions that are connected with different layers in the training progress. Second, attributes having complex localizing characteristics and different scales, the way only using features learned in the last layer is inappropriate, because the features learned in the last layer are too global to local attributes

(e.g., *has sunglasses*, *upperBodyVNeck* and *footwearSandals*). Therefore, as shown in Fig. 2, in the proposed MSMLCNN, supervisory signals are added by fully connecting each attribute with multiple pooling layers at different scales, which makes the MSMLCNN able to learn multi-scale features for different attributes and apply multi-scale features for the testing phase. Parameter details of the proposed MSMLCNN is listed in Table 1. In the table C, MP and AP represent convolutional, max pooling and average pooling layers, respectively.

Table 1 Parameter details of the proposed MSMLCNN

Layer	Type	Size	Neuron	Filter/stride
1	C	128×48×64	ReLU	3×3/1
2	C	128×48×64	ReLU	3×3/1
3	MP	64×24×64	-	3×3/2
4	C	64×24×96	ReLU	3×3/1
5	C	64×24×96	ReLU	3×3/1
6	MP	32×12×96	-	3×3/2
7	C	32×12×128	ReLU	3×3/1
8	MP	16×6×128	-	3×3/2
9	C	16×6×160	ReLU	3×3/1
10	MP	8×3×160	-	3×3/2
11	C	8×3×192	-	3×3/1
12	AP	6×1×192	-	3×3/1

2.2 Cost function design

In order to make the proposed MSMLCNN able to predict all attributes simultaneously, all attributes are fully connected with the last three pooling layers, as shown in Fig. 2. In practice, there are not only binary class attributes, but also multi-class attributes, such as clothing color. In order to make the design of each attribute's cost function more consistent, the pedestrian attribute classification problem is transformed into a multi-label classification problem including multiple binary attributes needing to be classified, then all binary attribute classification problems will be solved with logistic regression models[41]. The cost function of the proposed MSMLCNN is formulated as follows:

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

where G_k is the cost function of the k-th attribute, K is the total number of attributes, $\lambda_k \geqslant 0$ is a parameter used to control the contribution of the k-th attribute classification. In this work, λ_k is set as $\lambda_k = 1/K$ and G_k is formulated as follows:

$$G_k = -\frac{1}{N}\sum_{n=1}^{N}[y_k^n \log(h_{w_k}(x^n)) + (1-y_k^n)\log(1-h_{w_k}(x^n))]$$

$$h_{w_k}(x^n) = \frac{1}{1+e^{-w_k x^n}} \tag{2}$$

where $\{x_n, y_k^n\}$ represents a training sample and $y_k^n \in \{0,1\}$ is k-th attribute label of n-th sample x_n, N represents the number of training samples, w_k represents fully connected parameters between k-th attribute and the last pooling layers. To avoid the imbalanced classification, Eq. (2) is further extended as:

$$G_k = -\frac{1}{N}\sum_{n=1}^{N}[y_k^n \log(h_{w_k}(x^n))\beta_k^1 + (1-y_k^n)\log(1-h_{w_k}(x^n))\beta_k^0]$$

$$\beta_k^0 = \frac{N_k^1}{N},\ \beta_k^1 = \frac{N_k^0}{N} \tag{3}$$

where N_k^0 and N_k^1 are the numbers of negative and positive samples of k-th attribute, respectively. The learning toolbox for MSMLCNN is the cuda-convnet and it can be found in a google code website, https://code.google.com/p/cuda-convnet/.

3 Experiment and analysis

The challenging database PETA[11] is used to validate the superiority of the proposed MSMLCNN based pedestrian classification method. The PETA database consists of 10 subsets, such as VIPeR, PRID, GRID, CAVIAR4REID. Therefore, the PETA database is very complex which contains variations of different camera views, illuminations, resolutions and scenarios. PETA includes 19000 images and each image is annotated with 65 attribute, such as *gender*, *age*, *hair length* and *clothing color*.

A comparison is made for two baseline methods proposed in Ref. [42] and MSMLCNN. The first baseline method ikSVM[42] is a SVM-based method. The second method MRFr2[42] exploits the context of neighboring images by a Markov random field (MRF) to improve the performance. The MRF is an undirected graph, where each node represents a random variable and each edge represents the relation between two connected nodes. The unary energy item is the probability predicted by ikSVM, while the pairwise energy item is the similarity between two neighboring images learned by a random forest (RF) method. Both two baseline methods use foreground masks to improve feature extraction.

For the method, both multi-scale and single scale configurations are evaluated. As shown in Fig. 2, the multi-scale configuration of multi-label CNN (MSMLCNN) uses the features learned in the last three pooling layers (i.e., layers 8, 10, 12). The single scale configuration of multi-label CNN (SSMLCNN) only uses

the features learned in the last pooling layer (i. e., layer 12).

3.1 Setup

All images of the PETA database are scaled into 128×48 pixels. Following the evaluation protocol in Ref. [11], PETA is divided into non-overlapping train, validation and test subsets, which includes 9500, 1900, and 7600 images, respectively. Both train and validation subsets are augmented by translation and mirror operations. All multi-class attributes are transformed into multiple binary-class attributes. A binary attribute is considered imbalanced, if the number of positive samples is less than 50, and it is discarded. After discarding imbalanced attributes, 82 binary attributes were got, as shown in Table 2. For these 82 attributes, each attributes' classification accuracy and recall rate are reported when false positive rate (FPR) is set at 10%, along with the Area Under the ROC Curve (AUC), while the two baseline methods in Ref. [42] only provide 35 attributes' classification accuracies. For both the SSMLCNN and MSMLCNN, the average pooling layer (i. e., layer 12) is followed with a 0.5 ratio dropout layer. Moreover, both SSMLCNN and MSMLCNN use default thresholds (i. e., 0.5) to obtain the classification accuracy of each attribute.

3.2 Convergence comparison between SSMLCNN and MSMLCNN

As shown in Fig. 3, MSMLCNN converges faster than SSMLCNN on the train subset. Moreover, on the validation subset, MSMLCNN also obtains better performance than SSMLCNN. This result illustrates the way of adding supervisory signals by fully connecting each attribute with multiple pooling layers at different scales is helpful to train a deep CNN.

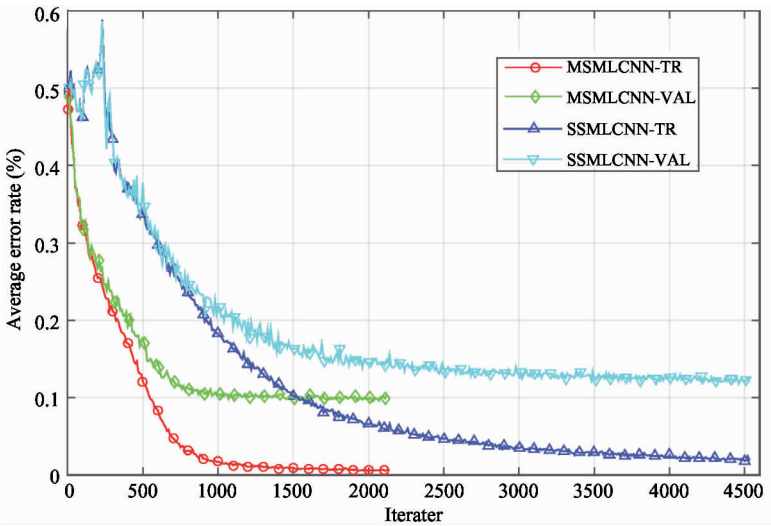

Fig. 3 The convergence comparison between SSMLCNN and MSMLCNN (TR and VAL represents the training and validation subsets)

3.3 Performance comparison

In Table 2, the results of the two baseline methods are directly cited from Ref. [42] and only 35 attributes' accuracies are obtained, thus for those results that were not reported in Ref. [42] are recorded as N/A. For the first 35 attributes, from Table 2, it can be found that both SSMLCNN and MSMLCNN excel the two baseline methods for most attributes and get 9.5% and 12.0% average accuracy improvements to the better baseline method (i. e., MRFr2), respectively.

Table 2 The performance comparison of ikSVM[42], MRFr2[42], SSMLCNN and MSMLCNN

Attribute	Accuracy rate (%)				Recall rate(%)@FPR=10%		AUC(%)	
	ikSVM	MRFr2	SSML-CNN	MSML-CNN	SSML-CNN	MSML-CNN	SSML-CNN	MSML-CNN
1. age16-30	83.1	86.8	75.2	79.1	48.7	60.0	82.01	86.32
age31-45	77.6	83.1	75.1	79.1	47.5	56.6	78.86	82.76
age46-60	79.1	80.1	90.3	91.7	58.6	64.9	79.55	85.00
ageabove60	93.5	93.8	95.7	97.3	82.6	88.7	92.46	94.78

Continued

5. backpack	70.7	70.5	78.3	82.3	44.5	49.6	80.23	82.83
carrying other	66.9	73.0	76.1	77.5	39.1	45.3	74.26	77.79
lower casual	76.5	78.2	87.0	90.2	30.3	42.1	80.16	84.64
upper casual	76.0	78.1	85.4	89.1	30.9	40.6	79.71	83.83
lower formal	76.6	79.0	87.5	90.2	57.7	67.5	82.33	86.48
10. upper formal	76.8	78.7	87.6	90.7	57.6	68.1	81.64	86.30
hat	89.4	90.4	92.8	95.0	77.2	82.6	89.15	91.53
upper jacket	69.6	72.2	89.7	93.4	49.8	54.8	80.74	81.95
lower jeans	79.8	81.0	77.5	82.1	52.2	65.7	80.83	86.64
leather shoes	84.0	87.2	80.2	83.5	60.2	68.5	83.37	88.43
15. upper logo	53.4	52.7	89.2	91.9	33.8	45.0	74.50	79.76
long hair	79.4	80.1	81.1	84.7	54.9	66.5	81.86	87.63
male	84.6	86.5	76.5	81.1	52.6	66.0	82.91	88.64
messenger bag	74.8	78.3	75.5	76.3	49.4	53.9	76.83	80.28
muffler	92.2	93.7	95.3	96.1	80.2	89.4	91.50	95.32
20. no carrying	72.5	76.5	75.9	78.6	45.9	49.7	76.69	80.32
no accessory	79.2	82.7	82.2	82.9	31.8	46.0	80.27	83.93
upper plaid	65.1	65.2	94.5	95.2	39.6	51.8	79.09	84.15
plastic bags	79.0	81.3	90.0	93.7	60.5	70.6	82.56	86.99
sandals	51.9	52.2	95.1	96.9	44.1	60.5	79.82	85.56
25. shoes	72.0	78.4	70.8	74.6	44.1	49.0	75.61	78.93
lower shorts	65.2	65.2	94.3	95.5	58.6	71.1	84.61	89.86
upper short sleeve	75.1	75.8	86.7	87.8	53.6	63.0	82.93	88.36
lower short skirt	69.6	69.6	93.6	94.3	60.2	66.9	84.25	87.65
upper thin stripes	51.9	51.9	93.0	96.8	28.2	23.7	71.33	70.92
30. upper sweater	71.5	75.0	94.8	96.6	53.4	53.0	80.36	78.74
sunglasses	53.3	53.5	95.0	96.0	46.8	60.7	82.07	87.01
lower trousers	77.9	82.2	70.6	74.7	39.9	52.2	77.49	82.64
upper T-shirt	71.1	71.4	90.5	91.3	49.7	61.5	81.25	87.36
upper v-neck	53.3	53.3	96.5	97.7	40.9	52.3	75.59	81.53
35. upper other	83.2	87.3	75.0	78.9	62.6	67.5	82.79	86.65
average(1-35)	73.6	75.6	85.1	87.6	50.1	59.2	80.80	85.02
ageless15	N/A	N/A	98.5	99.2	64.8	69.0	88.41	88.84
hair band	N/A	N/A	94.3	94.8	44.2	46.8	76.91	79.03
kerchief	N/A	N/A	99.5	99.6	85.7	87.0	92.60	93.22
baby buggy	N/A	N/A	97.9	99.1	80.0	90.9	91.90	96.36
40. folder	N/A	N/A	94.1	98.0	40.3	42.5	73.12	74.60
luggage case	N/A	N/A	96.0	98.3	66.9	79.9	88.25	93.14
suitcase	N/A	N/A	95.9	97.9	59.8	63.0	86.89	89.40
lower hot pants	N/A	N/A	98.8	99.5	89.6	90.3	95.71	95.48
lower capri	N/A	N/A	95.1	97.2	38.6	51.8	77.22	82.84
45. lower suits	N/A	N/A	94.4	95.6	59.7	73.7	84.51	89.53
lower long skirt	N/A	N/A	96.7	98.3	70.1	73.0	87.19	87.99
upper suit	N/A	N/A	95.2	96.3	59.6	73.5	84.33	89.01
upper long sleeve	N/A	N/A	86.3	88.0	33.8	63.2	81.23	87.33
upper no sleeve	N/A	N/A	96.6	97.9	69.9	74.7	87.68	92.44
50. upper thick stripes	N/A	N/A	96.5	98.5	28.3	45.0	70.52	78.81
short hair	N/A	N/A	78.4	83.3	38.9	59.7	80.30	86.14
hair bald	N/A	N/A	96.8	98.1	53.0	66.9	77.73	86.55

Continued

hair black	N/A	N/A	81.0	86.0	56.0	73.4	86.15	91.19
hair brown	N/A	N/A	86.5	88.8	66.3	75.3	85.92	89.85
55. hair grey	N/A	N/A	92.6	95.1	59.9	69.9	81.33	87.03
hair white	N/A	N/A	97.3	98.5	85.1	91.7	93.61	96.97
hair yellow	N/A	N/A	93.5	96.2	61.4	77.2	85.72	91.35
upper black	N/A	N/A	82.5	84.8	72.1	77.2	89.73	92.29
upper blue	N/A	N/A	91.9	94.9	71.4	80.7	88.31	92.80
60. upper brown	N/A	N/A	91.5	93.7	63.9	71.0	85.84	89.22
upper green	N/A	N/A	92.5	97.3	49.8	70.2	80.37	89.02
upper grey	N/A	N/A	81.4	85.1	47.2	53.0	80.27	83.07
upper purple	N/A	N/A	93.5	97.8	50.0	71.2	77.35	89.77
upper red	N/A	N/A	93.8	96.6	75.0	87.6	90.94	94.94
65. upper white	N/A	N/A	86.2	88.0	64.4	72.9	86.64	90.43
upper yellow	N/A	N/A	95.4	98.8	61.0	77.9	82.84	91.95
upper pink	N/A	N/A	96.7	98.3	47.8	70.4	82.15	88.39
upper orange	N/A	N/A	96.6	98.9	66.7	77.0	88.74	88.77
lower black	N/A	N/A	79.8	82.2	61.8	70.6	87.17	90.20
70. lower blue	N/A	N/A	84.5	87.0	65.3	77.9	86.55	91.45
lower brown	N/A	N/A	94.5	96.4	61.4	73.6	84.67	90.15
lower grey	N/A	N/A	76.8	82.9	43.6	55.7	76.56	83.23
lower red	N/A	N/A	98.4	99.2	72.4	82.8	90.73	95.44
lower white	N/A	N/A	93.6	94.6	68.8	78.2	87.93	91.71
75. footwearblack	N/A	N/A	70.6	74.6	44.7	56.4	77.65	83.08
footwear brown	N/A	N/A	89.0	91.2	55.6	70.0	81.33	87.42
footwear grey	N/A	N/A	83.7	84.6	44.3	49.2	77.53	79.98
footwear red	N/A	N/A	93.9	98.4	59.3	72.8	85.94	92.20
footwear white	N/A	N/A	79.7	82.8	40.0	57.9	77.91	84.30
80. boots	N/A	N/A	94.3	96.2	63.6	70.8	85.18	87.47
sneakers	N/A	N/A	78.1	78.9	39.9	49.2	78.79	83.00
stocking	N/A	N/A	95.3	96.3	73.3	78.2	89.53	91.65
average(1-82)	N/A	N/A	88.7	91.1	55.4	65.4	82.77	87.08

Comparing MSMLCNN with SSMLCNN, it can be observed that MSMLCNN obtains better performances for almost all attributes, thus higher average recall rate and AUC are obtained, as shown in Fig. 4 and Fig. 5. Specifically, for the first 35 attributes, the average recall rate and AUC of MSMLCNN are 9.1% and 4.22% higher than those of SSMLCNN, respectively. Moreover, for all 82 attributes, the average recall rate and AUC of MSMLCNN are 10.0% and 4.31% higher than those of SSMLCNN, respectively. The first 35 attributes' and 82 attributes' average ROC curves resulting from SSMLCNN and MSMLCNN are shown in Fig. 4 and Fig. 5, respectively. Both Fig. 4 and Fig. 5 show that MSMLCNN has better ROC performances than SSMLCNN. These results demonstrate that MSMLCNN learning multi-scale features is helpful for improving the pedestrian attribute classification performance.

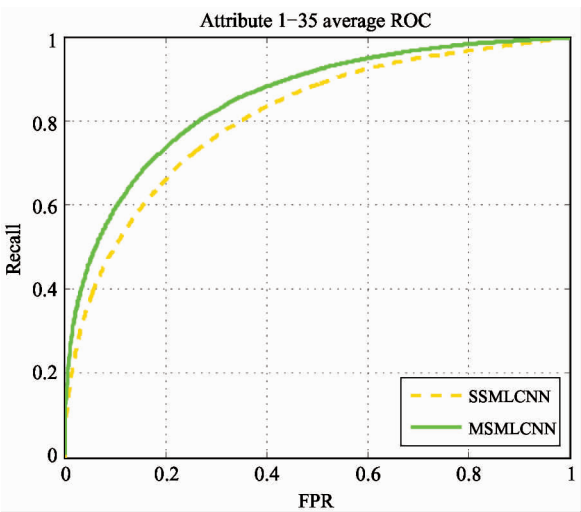

Fig. 4 The average (1-35) attribute ROC curves by using SSMLCNN and MSMLCNN

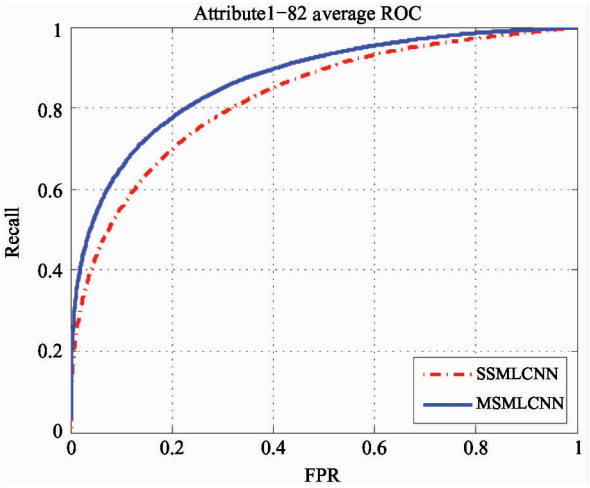

Fig. 5 The average (1-82) attribute ROC curves by using SSMLCNN and MSMLCNN

4 Conclusion

In this paper, a multi-scale and multi-label convolutional neural network (MSMLCNN) is proposed to predict multiple pedestrian attributes simultaneously. The multi-attribute classification problem is transformed into a multi-label classification problem including multiple binary attribute classification problems. Then, those multiple binary attribute classification problems are simultaneously solved by fully connecting each attribute with multi-scale features learned by the MSMLCNN. The multi-scale features are obtained by concatenating those featured maps produced from multiple pooling layers of the MSMLCNN at different scales. The way of using multi-scale features for pedestrian attribute classification has two benefits: avoiding gradient vanish to make the optimization of a very deep neural network easier; improving attribute classification accuracies by applying both local and global features. Extensive experiments show that proposed MSMLCNN outperforms state-of-the-art methods with a large margin.

-Reference

[1] Vaquero D, Feris R, Tran D, et al. Attribute-based people search in surveillance environments[C]. In: Proceedings of IEEE Workshop on Applications of Computer Vision, Snowbird, Utah, 2009. 1-8

[2] Jaha E S, Nixon M S. Analysing soft clothing biometrics for retrieval[C]. In: Proceedings of the 1st International Workshop on Biometric Authentication, Sofia, Bulgaria, 2014. 234-245

[3] Jaha E S, Nixon M S. Soft biometrics for subject identification using clothing attributes[C]. In: Proceedings of IEEE International Joint Conference on Biometrics, Clearwater, Florida, USA, 2014. 1-6

[4] Reid D A, Nixon M S, Stevenage S V. Soft biometrics; human identification using comparative descriptions[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(6):1216-1228

[5] Martinson E, Lawson W, Trafton J G. Identifying people with soft-biometrics at fleet week[C]. In: Proceedings of ACM/IEEE International Conference on Human-Robot Interaction, Chicago, USA, 2013. 49-56

[6] Layne R, Hospedales T M, Gong S G, et al. Person re-identification by attributes[C]. In: Proceedings of British Machine Vision Conference, Guildford, UK, 2012. 1-8

[7] Zhu J Q, Liao S C, Yi D, et al. Multi-label CNN based pedestrian attribute learning for soft biometrics[C]. In: Proceedings of International Conference on Biometrics, Phuket, Thailand, 2015. 535-540

[8] Zhu J Q, Liao S C, Lei Z, et al. Pedestrian attribute classification in surveillance: Database and evaluation [C]. In: Proceedings of Workshop on IEEE International Conference on Computer Vision, Sydney, Australia, 2013. 331-338

[9] Layne R, Hospedales T M, Gong S G. Attributes-based Re-identification[M]. London: Springer, 2014. 93-117

[10] An L, Chen X J, Kafai M, Yang S F, et al. Improving person re-identification by soft biometrics based re-ranking [C]. In: Proceedings of International Conference on Distributed Smart Cameras, Palm Springs, USA, 2013. 1-6

[11] Deng Y B, Luo P, Loy C C, et al. Pedestrian attribute recognition at far distance[C]. In: Proceedings of ACM Multimedia, Orlando, USA, 2014. 789-792

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. https://arxiv.org/abs/1409.1556: arxiv 2014

[13] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]. In: Proceedings of International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 2010. 249-256

[14] Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking[C]. In: Proceedings of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Rio de Janeiro, Brazil, 2007. 1-7

[15] Hirzer M, Beleznai C, Roth P M, et al. Person Re-identification by Descriptive and Discriminative Classification [M]. London: Springer, 2011. 91-102

[16] Liu C X, Gong S G, Loy C C, et al. Person re-identification: what features are important? [C]. In: Proceedings of Workshop on European Conference on Computer Vision, Florence, Italy, 2012. 391-401

[17] Zhang X Y. Simultaneous optimization for robust correlation estimation in partially observed social network[J]. *Neurocomputing*, 2016, 205: 455-462

[18] Zhu X B, Liu J, Wang J Q, et al. Sparse representation for robust abnormality detection in crowded scenes[J]. *Pattern Recognition*, 2014, 47(5):1791-1799

[19] Friedman J, Hastie T, Tibshirani R, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)[J]. *The Annals*

of Statistics, 2000, 28(2):337-407
[20] Zhang X Y, Wang S P, Zhu X B, et al. Update vs. upgrade: modeling with indeterminate multi-class active learning[J]. *Neurocomputing*, 2015, 162:163-170
[21] Chen H Z, Gallagher A, Girod B. Describing clothing by semantic attributes [C]. In: Proceedings of European Conference on Computer Vision, Florence, Italy, 2012. 609-623
[22] Zhu J Q, Liao S C, Lei Z, et al. Improve pedestrian attribute classification by weighted interactions from other attributes[C]. In: Proceedings of Workshop on Asian Conference on Computer Vision, Singapore, 2014. 545-557
[23] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11):2278-2324
[24] Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations[C]. In: Proceedings of Annual International Conference on Machine Learning, Montreal, Canada, 2009. 609-616
[25] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. In: Proceedings ofAdvances in Neural Information Processing Systems, Harrahs and Harveys, Lake Tahoe, USA, 2012. 1097-1105
[26] Donahue J, Jia Y Q, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition[EB/OL]. https://arxiv.org/abs/1310.1531: arxiv, 2013
[27] Razavian A S, Azizpour H, Sullivan J, et al. Cnn features off-the-shelf: an astounding baseline for recognition [C]. In: Proceedings of Workshop on IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014. 512-519
[28] Sun Y, Wang X G, Tang X O. Deep learning face representation from predicting 10,000 classes[C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,Columbus, USA, 2014. 1891-1898
[29] Sun Y, Chen Y H, Wang X G, et al. Deep learning face representation by joint identification-verification[C]. In: Proceedings of Advances in Neural Information Processing Systems, Montréal, Canada, 2014. 1988-1996
[30] Sun Y, Liang D, Wang X G, et al. Deepid3: Face recognition with very deep neural networks[EB/OL]. https://arxiv.org/abs/1502.00873: arxiv, 2015
[31] Gong Y C, Jia Y Q, Leung T, et al. Deep convolutional ranking for multi-label image annotation[EB/OL]. https://arxiv.org/abs/1312.4894: arxiv, 2013
[32] Zhu J Q, Liao S C, Lei Z, et al. Multi-label convolutional neural network based pedestrian attribute classification [J]. *Image & Vision Computing*, 2016, 58(C):224-229
[33] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines[C]. In: Proceedings of International Conference on Machine Learning, Haifa, Israel, 2010. 807-814
[34] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [EB/OL]. https://arxiv.org/abs/1207.0580: arxiv, 2012
[35] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [EB/OL]. https://arxiv.org/abs/1502.03167: arxiv, 2015
[36] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions [EB/OL]. https://arxiv.org/abs/1409.4842: arxiv, 2014
[37] Sermanet P, LeCun Y. Traffic sign recognition with multi-scale convolutional networks[C]. In: Proceedings of International Joint Conference on Neural Networks, Alaska, USA, 2011. 2809-2813
[38] He K M, Zhang Y Z, Ren S Q, et al. Deep residual learning for image recognition[EB/OL]. https://arxiv.org/abs/1512.03385: arxiv, 2015
[39] He K M, Zhang Y Z, Ren S Q, et al. Identity mappings in deep residual networks[C]. In: Proceedings of European Conference on Computer Vision, Amsterdam, Netherlands, 2016. 630-645
[40] Huang G, Liu Z, Maaten L V D, et al. Weinberger. Densely connected convolutional networks[C]. In: Proceedings of Computer Vision and Pattern Recognition, Honolulu, USA, 2017. 2261-2269
[41] Hosmer D W, Lemeshow S. Introduction to the Logistic Regression Model [M]. 2nd Edition. Hoboken, New Jersey, USA: John Wiley & Sons, 2000. 1-30
[42] Deng Y B, Luo P, Loy C C, et al. Learning to recognize pedestrian attribute [EB/OL]. https://arxiv.org/abs/1501.00901: arxiv, 2015

Zhu Jianqing, born in 1987. He received the Ph. D. degree in Institute of Automation, Chinese Academy of Sciences in 2015. He also received the B. S. degree in Communication Engineering and the M. S. degree in communication and Information System from the School of Information Science and Engineering, Huaqiao University in 2009 and 2012, respectively. He is currently an assistant professor at the College of Engineering, Huaqiao University, Quanzhou, China. His current research interests include computer vision and pattern recognition, with a focus on image and video analysis, particularly person re-identification, object detection and video surveillance. He was awarded the best biometrics student paper award at the International Conference on Biometrics in 2015.