PEDESTRIAN ATTRIBUTE RECOGNITION BASED ON MULTI-TASK DEEP LEARNING AND LABEL CORRELATION ANALYSIS

Zuhe LI[1], Mengze XUE[2], Qian SUN[3]*, Chenyang LIU[4], Qingbing GUO[5], Fengqin WANG[6], Lujuan DENG[7], Huanlong ZHANG[8]

Pedestrian attribute recognition is an extremely challenging assignment because of continual appearance variations, background clutter, pedestrian occlusion, and diverse spatial distribution of unbalanced attributes. Thus, we propose a multi-task deep model for pedestrian attribute recognition, which aims at the various attributes of each pedestrian and the poor quality of pedestrian images. In this model, a deep convolutional network called Mask R-CNN is firstly adopted to obtain binary masks of pedestrian bodies. Second, multiply by the features obtained from different convolutional layers and corresponding binary masks to eliminate background interference from the extracted image features. Then, select the most suitable combination of feature maps for each attribute by a voting mechanism. Finally, employ a correlation coefficient and conditional probability-based label analysis algorithm to integrate prior knowledge into the proposed network. This model can not only reduce the effects of image background, but also avoid the contradiction between recognition results of different attributes by establishing correlations between them. Our experiments are conducted on two datasets (RAP and PETA) with a large number of pedestrian images. Experimental results show that this method is superior to other existing methods.

Keywords: Pedestrian attribute recognition, Multi-task learning, Label correlation analysis

[1] Prof., School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China, e-mail: zuheli@126.com
[2] M. Eng., School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China, e-mail: xuemz158@163.com
[3] M. Eng., School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China, e-mail: 331907020384@zzuli.edu.cn
[4] M. Eng., School of Foreign Languages, Zhengzhou University of Light Industry, China, e-mail: chenyangliu1230@163.com
[5] M. Eng., School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China, e-mail: zzuli_gqb@163.com
[6] Prof., School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China, e-mail: wfq126@126.com
[7] Prof., School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China, e-mail: lujuandeng@163.com
[8] Prof., College of Electric and Information Engineering, Zhengzhou University of Light Industry, China, e-mail: zhl_lit@163.com

1. Introduction

In wake of the widespread presence of surveillance camera system and artificial intelligence, automatic identification of pedestrian attributes has become a long-term goal in the area of intelligent video analysis [1]. Pedestrian Attribute Recognition (PAR) purpose of digging for the attribute information of pedestrians across images, including age, gender, hair, clothing and so on. These attributes are widely used in various intelligent video analysis applications because of the plentiful semantic information included in them. For example, it is applied to the research of pedestrian re-recognition [2,3], person retrieval [4], and face verification [5]. However, there are still some handicaps in pedestrian attribute recognition for real video surveillance scenarios. For instance, occlusion, low resolution and continual appearance variations are still crucial problems need to be solved.

In the last few years, in wake of the improvement of deep learning framework in many computer-vision tasks, deep learning based PAR has become the mainstream. These deep learning methods for PAR can be generally classified into three types. (1) Dig out correlation between attributes: Probabilistic graphical model or adjacency matrices between tags are used to explore the correlation and mutual exclusion between pedestrian attributes [6,7]. For example, there is a positive correlation between the attributes of "female" and "long hair", but there is a mutually exclusive relationship between attributes corresponding to different age groups. Therefore, the correlation between these attributes can make effective constraint reasoning in multi-label classification tasks. (2) Employ attention mechanism: Li et al. [8] proposed an attention-based neural network, which mainly used the channel attention mechanism to extract and adjust the most pertinent and significant visual features of pedestrian attributes. Sarafianos et al. [9] proposed a portable network architecture for pedestrian attributes and a loss function for dealing with data imbalance. The network mainly adopts a multi-scale attention mechanism to target specific information of pedestrian attributes. Although the recognition accuracy of these methods had been improved, they did not consider the context of attributes. (3) Exploring visual context information: Ji et al. [10] proposed a PAR model based on CNN and LSTM neural network and explored the potential contextual relationship between attribute descriptions. Li et al. [11] proposed a new manner to identify human attributes, which further improve by exploring deeper contextual relationships from a human-centered perspective, we can distinguish between people and scene levels and reduce the influence of the environment. However, the existing PAR methods are using a single algorithm to explore the correlations between attributes, which ignore the impact factors that exist between different attributes, resulting in contradictions in the results of predicting attributes.

Dissimilar to the previous methods, we propose a multi-task depth model that do better in the areas of extracting foreground features, selecting an appropriate combination of feature maps and establishing the relationship between attributes completely and accurately. Specifically, the model composed of four parts. The first part is to use the backbone network to extract the local and global features of pedestrians and obtain four sets of feature layers of different scales. Second is to divide the image into two parts with the image segmentation model, which including pedestrian and background, then the binary mask image which correspond to the image multiply by different feature layers to eliminate the background [12]. Third is the construction of the voting mechanism. The most suitable feature combination for each attribute is selected so as to avoid losing too much image details by using only the highest feature layer to obtain attributes. Four is to integrate prior knowledge into the proposed network, utilizing correlation coefficients and label analysis algorithms based on conditional probability. The major contributions of this work can be generalized as follows:

- We propose a voting mechanism that effectively uses feature layers of different depths to obtain pedestrian attributes and choose the most appropriate feature layer for each attribute. It can effectively avoid the loss of fine-grained features and improve the performance of the model.
- We propose a label analysis algorithm that uses conditional probability and correlation coefficients to integrate prior knowledge into the network model. The algorithm is based on the principle of correlation and mutual exclusion between attributes.

2. Related work

In wake of the rapid development of the security industry, especially the vigorous advancement of construction projects such as Safe City, the demand for video surveillance platforms in finance, transportation and other fields continues to increase, and video surveillance is more and more widely used in the market. As the main object of video surveillance, reliable pedestrian attributes are extremely important. Therefore, many methods for detecting pedestrian attributes have emerged. Early pedestrian attribute recognition methods [13,14] relied on relatively single manual features, for instance color and texture histograms, which also combined with classification algorithms SVM and CRF. However, these methods in practical applications didn't work out very well. Recently, methods stem from deep learning frameworks have been widely used and achieved great success. Li et al. [15] proposed and compared two algorithms. One is to treat attribute recognition as multiple binary classification problems, and the other uses the correlation between attributes to build a model. It is proved by experiments that the correlation between attributes can make progress the performance of the

model. Wang et al. [16] proposed a combined loop (JRL) learning model to explore the context and relevance of attributes, which made progress RAP given small-scale training data with poor image quality. Chen et al. [17] proposed a novel attribute model stem from the relationship between pedestrian attributes and body parts, and combined low-level and high-level features to design a neural network to overcome the problem of pedestrian pose changes in video surveillance. Yan et al. [18] proposed a multi-task lightweight convolutional neural network model to realize pedestrian attribute recognition. Although these techniques adopted deep learning frameworks and taken the spatial and semantic relationships between attributes to further make the recognition performance of the model into consideration better, these methods required manual definition of rules, such as attribute groups and prediction sequences, which was difficult to determine in practice. Therefore, some researchers introduced the attention mechanism. Zeng et al. [19] proposed a novel Collaborative Attention Sharing (CAS) module that increases the exchange of spatial information, using different channels for feature fusion between tasks to generate attention and enhance task-specific features. Zhao et al. [20] proposed to use the combination of end-to-end cyclic convolution (RC) and cyclic attention (RA) models to partition all pedestrian attributes and mine the spatial locality and semantic correlations that exist between different attribute groups. Tang et al. [21] combined the features of multiple different convolutional layers by introducing the feature gold tower structure, learned the regional features of each attribute on multiple levels, and adaptively locates the information region of each attribute in a weakly supervised mode. Lin et al. [22] proposed a dynamic adaptive weight loss model that removed multiple branches in the training process and improved the recognition efficiency and accuracy.

3. Proposed method

Our main purpose is to build a feature pyramid structure and introduce a voting mechanism while eliminating background effects. At the same time, the correlation coefficient matrix and the conditional probability matrix are adopted to explore the correlation and mutual exclusion between attributes, avoiding problems such as unbalanced samples. This will help improve performance, especially when the image is not clear enough.

3.1. Overall structure

A conspectus of the proposed model is present in Fig. 1. In the single-shot example, complete attributes of the pedestrian are inferred from multiple tasks. Our model consists of 4 parts: 1) Backbone network module, which acts as a pedestrian feature extractor; 2) Body segmentation module, which uses the Mask R-CNN network to generate binary mask images, which is multiplied by 4 sets of

features with different depths after processing. Obtain 4 sets of foreground features; 3) Vote module, which will compare and select 4 different prediction results; 4) Correlate module, which based on attribute correlation coefficient matrix and conditional probability matrix. It will establish a pedestrian attribute selection mechanism, taking into account the relevance and mutual exclusion of pedestrian attributes.

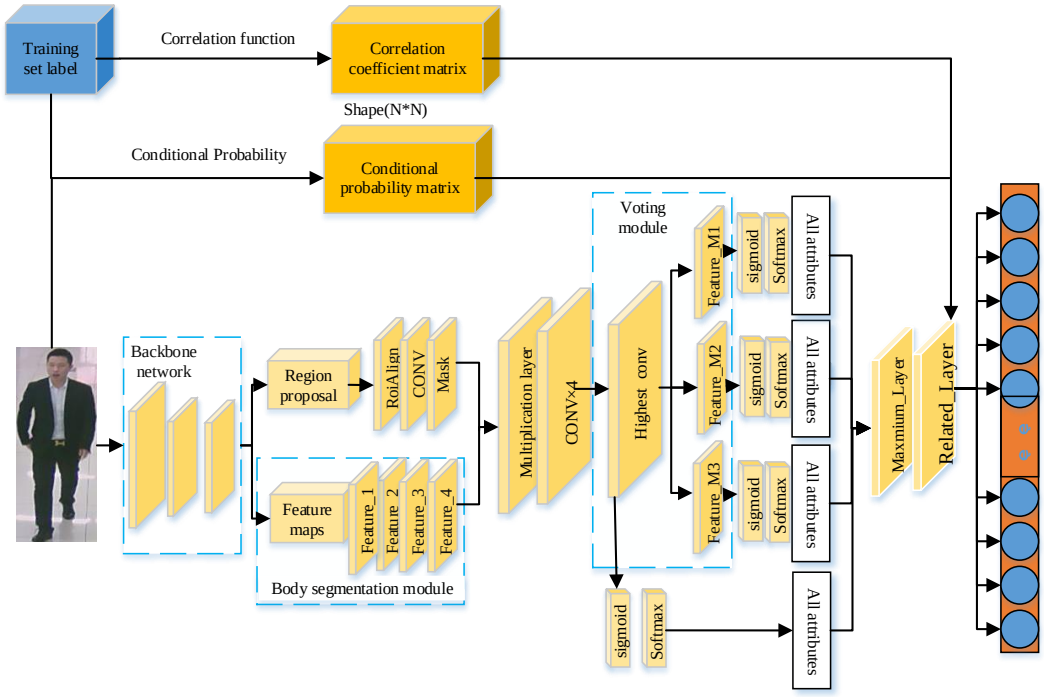

Fig. 1. The multi-task model used to feature extraction and image segmentation of input images. The feature pyramid is divided into four different feature layers. Each feature layer multiplied the binary mask image. Conv×4 represents 4 different convolutional layers. After Conv×4, features of different dimensions can be obtained. Then each layer selects sigmoid and softmax functions to predict all attributes of pedestrians. The maximum value of all attribute predictions can be get through the Maxmium_Layer layer, and the final pedestrian attributes are predicted through the Related_Layer layer according to the attribute relationship obtained from the correlation coefficient matrix and the conditional probability matrix.

First, images are sent to a set of convolutional layers to extract local and global features. Combined with pedestrian attributes, convolutional layers are divided into four groups of feature layers because of the features extracted by the convolutional layers of different depths are different. Next, pedestrian image be sent as a binary image by the human body segmentation module. The image background pixel is set to 0, and the pedestrian body pixel is set to 1. The background elements are removed by multiplying four groups of different feature

layers with the image background element by element. Then the feature layer with the background elements removed is convolved to detect all pedestrian attributes. Finally, for each task, we combine the results of the two matrices and use different activation functions to achieve mutual exclusion and correlation between attributes. Thus, the accuracy of network detection can be improved.

3.2. Backbone network

The backbone network part of the model uses ResNet50 (deep residual network) [23,24] to extract image features. The advantage of the ResNet network is that it proposes a residual structure, which not only deepens the convolution depth to obtain a larger receptive field, but also effectively avoids problems such as gradient explosion.

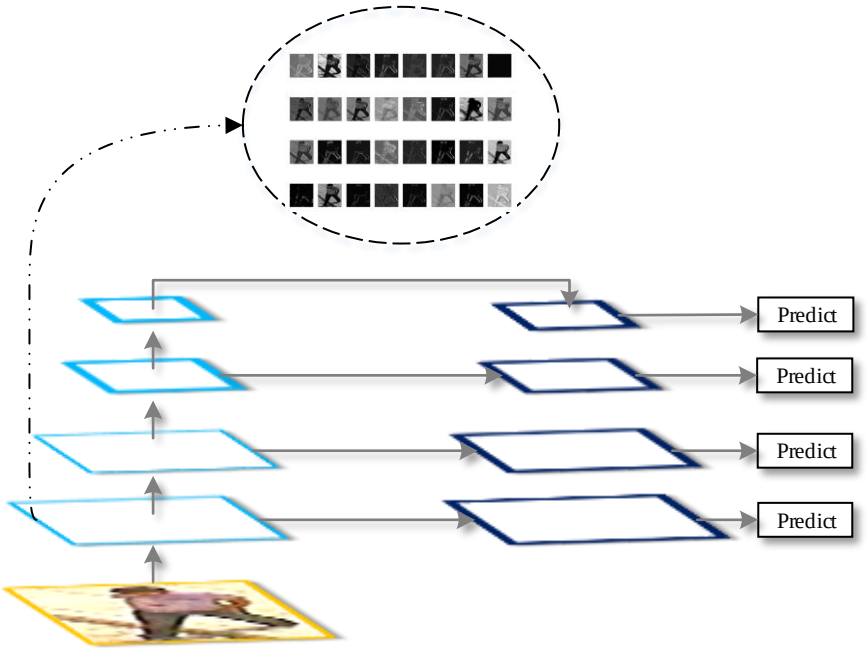

Fig. 2. Feature Pyramid Network. The dotted box in the figure shows a sample of one of the feature layers of the feature pyramid structure.

As we all know, the features in the deeper convolutional layer will lose some fine-grained features, as well as some finer details. If only the highest-level convolutional layer is used to identify pedestrian attributes, accurate recognition cannot be achieved for some pedestrian attributes that require fine-grained features, such as "glasses" and "shoes". Therefore, we recommend using the feature pyramid structure based on the ResNet50 network model. As shown in Fig. 2, we extracted four sets of feature layers with different depths, and predicted

all attributes at the same time, using a voting mechanism to build a certain correlation between the prediction results of the four sets of feature layers. The prediction result of each feature layer is different, and the result of each attribute is voted by four groups of feature layers. We assume that there are two cases, that is, when three or more of the four sets of outcomes are both greater or less than the set threshold, the maximum or minimum is selected. Otherwise, the prediction results of the highest convolution layer are used.

3.3. Foreground segmentation module

We use the Mask R-CNN [25,26] network model to acquire human segmentation images. The Mask R-CNN network uses the features acquired by the backbone network as the effective feature layer of a Region Proposal Network (RPN) network. The role of the RPN network is to determine whether the pixel contains an object, thereby generating proposal box. Then we use the mask model and classifier to decode and classify the proposal box. Given the input image *I*, multiply by the obtained segmented image with the four groups of different feature layers described above, element by element, to eliminate the background influence of each feature layer, which refer to equations 1:

$$G_i = \phi(I) \odot \varphi(I), i \in \{1,2,3,4\} \tag{1}$$

Where $\varphi_i(I)$ represents 4 groups of specific feature layers, $i \in \{1,2,3,4\}$, $\emptyset(I)$ represents a segmented image. Among them, $\varphi_i(I) \in R^{H_i \times W_i \times S_i}$), $\emptyset(I) \in R^{H_i \times W_i \times S_i}$, $\odot$ stands for element-wise multiplication.

3.4. Relationship between multitasking attributes

At present, the known pedestrian attribute recognition methods are more concerned with the correlation between attributes and not the mutual exclusion between attributes. Therefore, we propose to construct a conditional probability matrix C used by the potential correlation between attributes, as shown in formula:

$$[\, N_1 \; N_2 \; \ldots N_{a-1} \; N_a \,] \tag{2}$$

where N_a denotes the occurrence numbers of each attribute in the training set, then we count the co-occurrence of attribute pairs in the training set and get the matrix M:

$$\begin{bmatrix} \mathrm{M}_{11} & \mathrm{M}_{12} & \cdots & \mathrm{M}_{1,a-1} & \mathrm{M}_{1a} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ M_{a1} & \mathrm{M}_{a2} & \cdots & \mathrm{M}_{a,a-1} & \mathrm{M}_{aa} \end{bmatrix} \tag{3}$$

where M_{ab} denotes the co-occurrence times of $Label_a$ and $Label_b$. Then, we can get the conditional probability matrix by using formula $\mathrm{P}_{ab} = \mathrm{M}_{ab}/N_a$:

$$\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1,a-1} & P_{1a} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ P_{a1} & P_{a2} & \cdots & P_{a,a-1} & P_{aa} \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1,a-1} & M_{1a} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ M_{a1} & M_{a2} & \cdots & M_{a,a-1} & M_{aa} \end{bmatrix} / [\, N_1 \; N_2 \; \ldots N_{a-1} \; N_a \,] \quad (4)$$

As shown in formula 4: each row in P represents the probability value of one attribute and each of the other attributes at the same time, N_{ab} represents the conditional probability between attribute *a* and attribute *b*. Correlation is divided into weak correlation in the light of the size of the correlation coefficient; low correlation; significant correlation; high correlation. As shown in formula 6, it can be seen that the definition domain of the correlation coefficient is [-1.0,1.0], *a* and *b* respectively represent different attributes, and γ represents the correlation coefficient, where weak correlation is $|\gamma|<0.3$; low correlation is $0.3<|\gamma|<0.5$; Significantly weak correlation is $0.5<|\gamma|<0.8$; High correlation is $0.8<|\gamma|<1.0$ [27]. We constructed the correlation coefficient matrix between attributes according to formula 6, as shown in Fig. 3. Through Fig. 3, we observe that the true labels of the two attributes do not satisfy the mutually exclusive relationship even though the conditional probability value of one attribute and another attribute is 0.0. For example: {"Male", "Long Hair"}, {"Male", "Lb-Tight Trousers"}, these attributes themselves have a lower probability of existence than other attributes, which cannot be accurately represented by conditional probability alone the relationship between attributes. Therefore, the conditional probability matrix and the correlation coefficient matrix need to be combined. What should we do in the first place is selecting the attribute group whose conditional probability value between the two attributes is greater than or equal to 0.7 or equal to 0.0, then making further judgments derived from the correlation coefficient value between the attributes. The attribute relationship that satisfies the significant correlation and the high correlation is divided into positive effects; negative effects; mutual exclusion. As shown in Table1:

$$\gamma\,(a,b) = \frac{Cov(a,b)}{\sqrt{Var[a]Var[b]}} \quad (5)$$

Table 1

Partial attribute relationship

Mutually exclusive-attribute		
Male	0.0	Female
Age31-45	0.0	Age17-30
Body Normal	0.0	Body Fat
Black Hair	0.0	Hs- Hat
Employee	0.0	Customer
Lb-Jeans	0.0	Lb-Long Trousers
Positive impact		
Female	0.9925	Hs-Long Hair

Customer Hs-Black Hair Male	0.7804 0.9413 0.9403	Body Normal Customer Hs-Black Hair
Negative impact		
Long-Hair Lb-Tight Trousers Lb-Jeans Male Lb-skirt	0.0 0.0 0.0 0.0 0.0	Male Male Lb-Long Trousers Shoes-Boots Male

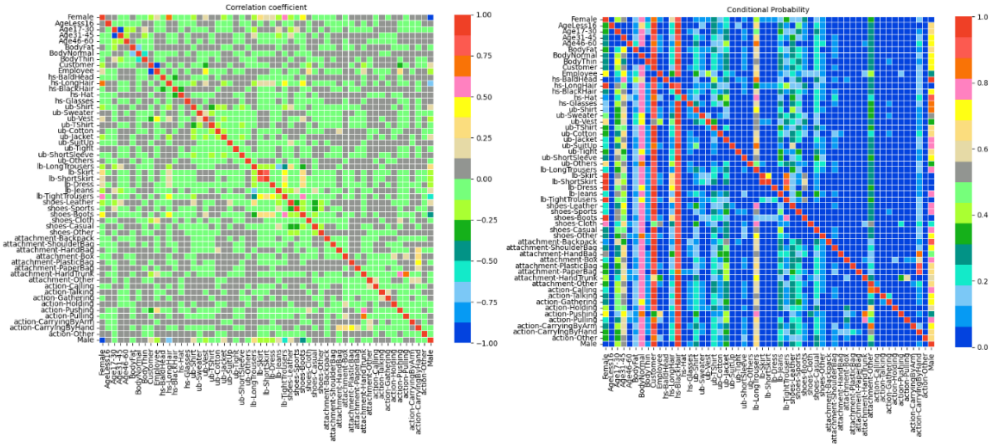

Fig. 3. Correlation coefficient and Conditional Probability

We propose to use different activation functions and an intervention mechanism at the same time to enhance the correlation and mutual exclusion between specific attributes. For mutually exclusive property, we choose to use the softmax activation function to get a set of two-dimensional vectors, which correspond to two mutually exclusive labels. For example, {"female", "male"}. For related properties, we use the sigmoid activation function to get a one-dimensional vector with a domain of [0.0, 1.0]. The intervention mechanism refers to through the use of the properties between prior knowledge and attributes and differential threshold before the change of the absolute value of the original forecast result. For example, {female, long hair}, {male, black hair}. P_f corresponds to the predicted label of women, P_m corresponds to the predicted label of men, P_l corresponds to the predicted label of long hair, P_h. corresponds to the predicted label of black hair, and the threshold K is 0.5. When P_f is greater than K and P_m is less than K, it is judged as female. When P_m is greater than K and P_f is less than K, it is judged as male. If the absolute numerical of the difference between P_f and K is greater than the absolute value of the difference between P_l and K, set P_l to P_f, and the determination methods of P_m and P_h are

the same as the former, all depending on the absolute value of the difference with the threshold K. As shown in formula 6:

$$(P_f, P_l)\begin{cases}\{Maxmium(|P_f - K|, |P_l - K|)\} = |P_f - K|, & P_l = P_f \\ \{Maxmium(|P_f - K|, |P_l - K|)\} = |P_l - K|, & P_f = P_l\end{cases} \tag{6}$$

When attribute relationships that have a negative influence on each other appear at the same time, such as {"male", "long hair"}, we refer to the absolute value of the difference with the threshold K. If the difference between P_m and K is greater than the difference between P_l and K, and is greater than the value of K, we set P_l to a value less than the value of K, and vice versa. As shown in formula :

$$(P_m, P_l)\begin{cases}\{Maxmium(|P_m - K|, |P_l - K|)\} = |P_m - K| \,\&\&\, (P_m, P_l) >= K, & P_l < K \\ \{Maxmium(|P_m - K|, |P_l - K|)\} = |P_l - K| \,\&\&\, (P_m, P_l) <= K, & P_m > K\end{cases} \tag{7}$$

For some action attributes with less relevance and mutual exclusion, we use the voting mechanism described above to get the prediction result. Use feature layers of different depths to make selection decisions through the Maxmium_layer layer to obtain the largest prediction result, as shown in formula 8:

$$O_t = \{Maxmium(\bar{O}_t, \tilde{O}_t)\} \tag{8}$$

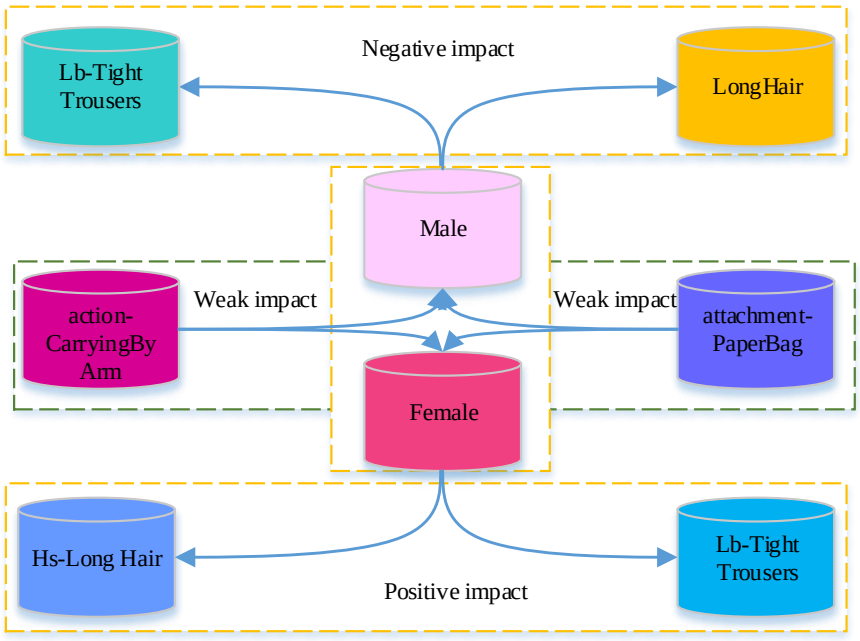

Fig. 4. attribute relationship

Among them, $\bar{O}_t$ represents the prediction result after the first three feature layers are compared with each other, $\tilde{O}_t$ represents the result of the most advanced convolutional layer detection, O_t represents the final prediction label, and *t* represents the category of the action attribute. The relationship between some attributes, as shown in Fig. 4:

The key to deep learning is the loss function. For the multi-classification problem of pedestrian attribute recognition, if the n-th image I_n, (n=1,...,N) has the a-th attribute, (a=1,...,A), then predict the image attribute label I_{na}=1; otherwise, I_{na}=0. The real label of the image is Y_{na}. The prediction function has the form $\Phi = \{\Phi_1, \Phi_2, \ldots, \Phi_A\}$, $\Phi_A(\mathrm{I}) \in \{1,0\}$. We define the minimization of the loss function over the training samples for the a-th attribute as:

$$\Psi_a = \arg\min_{\Psi_a} \sum_{n=1}^{N} L(\Phi_a(I_n, \Psi_a), Y_{na}) \tag{9}$$

Among them, Ψ includes a series of optimization parameters connected with the a-th attribute, and Φ_a comes back the predicted label of the a-th attribute of the n-th image. In addition, *Loss* is the loss function, which used to express the degree of the gap between the forecast and the actual data and to measure the quality of the model prediction.

For each pedestrian attribute using the sigmoid activation function, the loss function as shown in formula 10:

$$Loss = -\sum_{n=1}^{N} (Y_{na} \log(\vec{Y}_{na}) + (1 - Y_{na})(\log 1 - \vec{Y}_{na})) \tag{10}$$

Where Y_{na} represents the true label of a certain pedestrian image attribute, and $\vec{Y}_{na}$ represents the predicted label after using the sigmoid activation function.

For each pedestrian attribute that uses the softmax activation function, the loss function as shown in formula 11:

$$Loss = -\sum_{n=1}^{N} (S_{na} \ln(\vec{S}_{na})) \tag{11}$$

Where S_{na} represents the true label of a certain pedestrian image attribute, and $\vec{S}_{na}$ represents the predicted label after using the softmax activation function.

4. Experiments

Two well-known datasets are selected for evaluation: the PETA and the Richly Annotated Pedestrian (RAP), both of which are commonly used benchmarks in pedestrian attribute recognition experiments.

4.1. Datasets

The PETA dataset is made up of 19000 images and contains 10 different pedestrian image collections in outdoor environments. Each image contains sixty-one binary attributes and four multi-category attributes, totaling sixty-five attributes. This dataset selects different scenes and small-scale data sets under different conditions and takes images from various lighting scenes and various camera angles. The resolution of the image ranges from 17×39 to 169×365.

The RAP dataset is collected from real video scenes and contains 41585 images collected by 26 indoor surveillance cameras. Each pedestrian contains 72 fine-grained attributes, containing 69 binary attributes and 3 multi-value attributes, as well as some attributes such as shooting angle, occlusion, and body part information. This data set highlights environmental and contextual factors.

4.2. Implementation details

Our method is achieved by making use of Keras 2.3.1 and Tensorflow 1.14.0 backend. All experiments are performed on CPU E5-2680, 2.40GHz processor, TU102 [TITAN RTX] GPU. For the PETA dataset, we take 9500 samples as the training set, and set the batch size to 10 during training. After 46 epochs of training, we get the model with best performance, and the average time cost for each epoch is 838 seconds. For the RAP dataset, we take 20790 samples as the training set, and set the batch size to 15 during training. After 54 epochs of training, we get the model with best performance, and the average time cost for each epoch is 13184 seconds.

4.3. Comparison to other methods

In this section, we compare the performance obtained by our method with a number of most advanced methods. These methods are divided into three types: (1) Overall methods, comprising of ACN [28] method, which using only input images to jointly train all attributes without depending on external information. DeepMar [15] is an end-to-end CNN model that utilizes the relationship between attributes to identify all attributes at one time. MLCNN [7] realizes segmentation, detection and aggregation of input pedestrian images according to body parts. These methods are similar to ours that based on the global Method, using CNN network to extract the characteristics of the overall image of pedestrians. (2) Methods based on attribute relations, including SCRL [29] and Super-fine [30], which all use semantic relations through CNN-RNN-based models. (3) Attention-based method that can provide support to attribute localization module [21].

Table 2

Comparison between the results observed in the PETA dataset

Attributes	MLCNN [7]	DeepMar [15]	Proposed
AgeLess-30	81.1	85.8	86.2
AgeLess-45	79.9	81.8	83.4
AgeLess-60	92.8	86.3	94.1
AgeLarger-60	97.6	94.8	95.4
Long-Hair	88.1	88.9	87.9
Casual	89.3	84.4	92.0
Formal	91.1	85.1	93.2
Jacket	92.3	79.2	93.4
Short-Sleeves	88.1	87.5	91.3
T-shirt	90.6	83.0	91.8
Casual	90.5	84.9	92.7
Formal	90.9	85.2	92.8
Jeans	83.1	85.7	80.8
Trousers	76.2	84.3	88.4
Leather	85.2	87.3	89.8
Sneaker	81.8	78.7	83.2
Backpack	84.3	82.6	82.2
Carrying-Other	80.9	77.3	80.4

Table 2 displays the assessment results of DeepMar, MLCNN and our method on the PETA dataset. In light of the table, our model shows the excellent recognition rate of the main 18 (35 in total) attributes, and the overall accuracy rate has increased by more than 3%. The proposed network implements a recognition rate of 88.75%, while the value of the DeepMar method is 82.6% when 35 attributes are taken into account.

Table 3

Comparison between the results observed in the RAP dataset (average accuracy percentage)

Attributes	DeepMar [15]	ACN [28]	Proposed
Female	96.53	94.06	75.06
AgeLess-16	77.24	77.29	99.22
Age17–30	69.66	69.18	69.29
Age31–45	66.64	66.80	65.71
Age46–60	59.90	52.16	96.79
Body-Fat	61.85	58.42	86.31
Body-Normal	58.47	55.36	77.71
Body-Thin	55.75	52.31	92.60
Customer	82.30	80.85	94.08
Employee	85.73	85.60	94.88
Bald-Head	80.93	65.28	99.57
Long-Hair	92.47	89.34	91.53
Black-Hair	79.33	66.19	93.54

Hat	84.00	60.73	98.49
Sweater	64.21	56.85	92.57
Vest	89.91	83.65	95.47
T-shirt	75.94	71.61	76.67
Cotton	79.02	74.67	89.18
Jacket	80.69	78.29	79.56
Others	54.82	50.35	97.71
Long-Trousers	86.64	86.60	68.44
Skirt	74.83	70.51	97.36
Short-Skirt	72.86	73.16	98.11
Dress	76.30	72.89	97.07
Shoes-Boots	91.37	85.03	92.86
Backpack	80.61	68.87	97.92
Shoulder-Bag	82.52	69.30	93.16
Hand-Bag	76.45	63.95	97.50
Box	76.18	66.72	96.13
Other	58.79	54.83	99.15

Table 3 displays the assessment results of DeepMar, ACN and our method on the RAP dataset. In light of the table, the recognition rate of our model for most attributes has been improved. Among them, the average accuracy of ACN method prediction is about 68.9%, DeepMar method prediction is 75.5% and our method prediction is 90.87%. It is unreliable that accuracy of both female and male attributes can achieve 90% as the actual situation is not considered. The correlation coefficient matrix introduced in our experiment shows that women and men are completely mutually exclusive attributes, but in the prediction results of the comparison method, men and women exist at the same time, which lead to a higher accuracy rate. The correct experimental results are shown in Table 4:

Table 4

Give an example to show our prediction results

Gender	True_Label	Predicd_label	Acc_number	Accuracy
Female	(1,0,1,1,1,0,1)	(0,1,1,1,1,0,0)	4	57.14%
Male	(0,1,0,0,0,1,0)	(1,0,0,0,0,1,1)	4	57.14%

Table 5 lists the average accuracy of attributes of the same category, and the average accuracy of 8 category attributes. Among them, "L.hair" means hair length; "T.up" means top, "T.low" means lower body clothes and "T.shoes" means shoes. As can be seen, it has an average accuracy rate of 88.9% through our way. Compared with the SCRL method, the average accuracy rate is 0.9% lower because of the gender attribute. Few other methods can compare with ours in accuracy.

Table 5

Accuracy of attribute recognition

Method	age	gender	carrying	T.up	L.hair	accessory	T.low	T.shoes	Avg
SCRL[29]	89.5	92.1	88.2	88.3	92.3	93.8	86.7	86.9	89.8
Super-fine[30]	86.8	93.1	81.9	77.3	89.8	83.5	83.7	77.3	84.2
ALM[21]	-	-	-	-	-	-	-	-	79.5
Ours	89.7	77.3	90.1	92.3	92.0	94.2	87.3	88.6	88.9

4.5. Ablation experiment

In this section, we mainly carry out ablation experiments on the improved method we propose. Table 6 shows that when the voting mechanism is removed and the correlation between attributes is not considered, the accuracy rate of pedestrian attributes predicted by the RAP dataset model is 86.76%, and the accuracy rate of pedestrian attributes predicted by the PETA dataset model is 82.37%. Incorporating the voting mechanism (Voting Mechanism), the accuracy of the PETA dataset has increased by 1.61%, and the accuracy of the RAP dataset has increased by 1.02%. We believe that only using the top layer of the convolutional layer will result in loss of information. What's more, some attributes that require fine-grained features we did not consider. Finally, the result of combining the correlation coefficient matrix and the conditional probability matrix is used as the criterion for considering the correlation between attributes. Compared with only using the ResNet network, the accuracy of the PETA dataset is increased by 6.38%, and the RAP dataset is increased by 4.11%. When predicting the attributes of pedestrians, it will be affected by different attributes at the same time.

Table 6

Ablation experiment

Subset	Models	Precision	Accuracy
PETA	ResNet	79.20	82.37
	ResNet+VM	78.92	83.98
	VM+Correlation	78.76	88.75
RAP	ResNet	79.61	86.76
	ResNet+VM	77.84	87.78
	VM+Correlation	75.95	90.87

We list two sets of pictures. The left image is from the PETA dataset, and the right is from the RAP dataset. The two images are the original image and the binary mask image. These two images are used to predict pedestrian's attributes through our multi-task network model. As shown in Fig. 5:

Fig. 5. Pedestrian attribute prediction results

5. Conclusions

Background chaos, viewpoint changes and occlusion in video surveillance scenes have obvious negative effects on the performance of PAR methods. Therefore, this paper proposes a deep learning framework. Different feature layers are multiplied by binary masks images after a feature pyramid structure is implemented, while reducing the influence of the background. It enhances attribute positioning and regional feature-based learning, taking the principle of correlation and mutual exclusion between attributes into consideration, adding correlation coefficient matrix and conditional probability matrix to improve detection performance through intervention mechanism. Experimental results on PETA and RAP datasets demonstrate that the performance of this method is markedly better than most existing methods. The method proposed in this paper leads to a slow training speed by constructing a complex network model. In the future research, consider simplifying the training model to improve the performance and speed up the training efficiency.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61702462, 62072416, 61873246 and 61771432, the Scientific and Technological Project of Henan Province under Grant 222102210010 and 192102210108, and the Research and Practice Project of Higher Education Teaching Reform in Henan Province under Grant 2019SJGLX320 and 2019SJGLX020.

REFERENCES

[1]. *H.R.Yan, J.Y.Zhang, F.Li, T.Zhang, N.Li, Z.N.Li*, "Multi-Level Based Pedestrian Attribute Recognition", in 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing. IEEE, 2019, pp.166-169

[2]. *X.Ke, X.Lin, L.Qin*, "Lightweight convolutional neural network-based pedestrian detection and re-identification in multiple scenarios", in Machine Vision and Applications, **vol. 32**, no. 2, 2021, pp. 1-23

[3]. *S.Zhang, D.Chen, J.Yang, S.Bernt*, "Guided Attention in CNNs for Occluded Pedestrian Detection and Re-identification", in International Journal of Computer Vision, **vol. 129**, no. 6, 2021, pp. 1875-1892

[4]. *J.Wu, Y.Zhao, X.Liu*, "Enhancing person retrieval with joint person detection, attribute learning, and identification", in Pacific Rim Conference on Multimedia, Springer, Cham, 2018, pp. 113-124

[5]. *Y.X.Zhou, H.J.Ni, F.J.Ren, X.Kang*, "Face and gender recognition system based on convolutional neural networks", in 2019 IEEE International Conference on Mechatronics and Automation (ICMA), IEEE, 2019, pp. 1091-1095

[6]. *X.P.Song, H.B.Yang, C.C.Zhou*, "Pedestrian Attribute Recognition with Graph Convolutional Network in Surveillance Scenarios", in Future Internet, **vol. 11**, no. 11, 2019, pp. 245

[7]. *J.Q.Zhu, S.C.Liao, Z.Lei, S.Z.Li*, "Multi-label convolutional neural network based pedestrian attribute classification", in Image and Vision Computing, 2017, pp. 224-229

[8]. *Y.Li, H.H.Xu, M.J.Bian, J.S.Xiao*. "Attention based CNN-ConvLSTM for pedestrian attribute recognition", in Sensors, **vol. 20**, no. 3, 2020, pp. 811

[9]. *N.Sarafianos, X.Xu, I.A.Kakadiaris*, "Deep imbalanced attribute classification using visual attention aggregation", in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 680-697

[10]. *Z.Ji, W.Zheng, Y.Pang*, "Deep pedestrian attribute recognition based on LSTM", in 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 151-155

[11]. *Y.N.Li, C.Huang, C.C.Loy, X.O.Tang*, "Human attribute recognition by deep hierarchical contexts", in European Conference on Computer Vision. Springer, Cham, 2016, pp. 684-700

[12]. *E.Yaghoubi, D.Borza, J.Neves, A.Kumar, H.Poenca*, "An attention-based deep learning model for multiple pedestrian attributes recognition", in Image and Vision Computing, **vol. 102**, October, 2020, pp. 103981

[13]. *Y.B.Deng, P.Luo, C.C.Loy, X.O.Tang*, "Pedestrian attribute recognition at far distance", in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 789-792

[14]. *J.Zhu, S.Liao, Z.Lei, D.Yi, S.Li*, "Pedestrian attribute classification in surveillance: Database and evaluation", in Proceedings of the IEEE international conference on computer vision workshops, 2013, pp. 331-338

[15]. *D.W.Li, X.T.Chen, K.Q.Huang*, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios", in 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 111-115

[16]. *J.Wang, X.Zhu, S.Gong, W.Li*, "Attribute recognition by joint recurrent learning of context and correlation", in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 531-540

[17]. *Y.Q.Chen, S.Duffner, A.Stoian, J.Y.Dufour, A.Baskurt*, "Pedestrian attribute recognition with part-based CNN and combined feature representations", in VISAPP, 2018

[18]. *P.Yan, L.Zhuo, J.F.Li, H.Zhang, J.Zhang*, "Pedestrian Attributes Recognition in Surveillance Scenarios Using Multi-Task Lightweight Convolutional Neural Network", in Applied Sciences, **vol. 9**, no. 19, 2019, pp. 4182

[19]. *H.T.Zeng, H.Z.Ai, Z.J.Zhuang, L.Chen*, "Multi-task learning via co-attentive sharing for pedestrian attribute recognition", 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1-6

[20]. *X.Zhao, L.F.Sang, G.G.Ding, J.G.Han, N.Di, C.G.Yan*, "Recurrent attention model for pedestrian attribute recognition", in Proceedings of the AAAI Conference on Artificial Intelligence, **vol. 33**, no. 01, 2019, pp. 9275-9282

[21]. *C.F.Tang, L.Sheng, Z.X.Zhang, X.L.Hu*, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization", in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4997-5006

[22]. *X.Lin*, "Pedestrian Attribute Recognition Model based on Adaptive Weight and Depthwise Separable Convolutions", in 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), IEEE, 2020, pp. 830-833

[23]. *G.C.Porușniuc, F.Leon, R.Timofte, C.Miron*, "Convolutional neural networks architectures for facial expression recognition", in 2019 E-Health and Bioengineering Conference (EHB), IEEE, 2019

[24]. *K.M.He, X.Y.Zhang, S.Q.Ren, J.Sun*, "Deep residual learning for image recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778

[25]. *K.M.He, G.Gkioxari, P.Dollar, R.Girshick*, "Mask r-cnn", in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961-2969

[26]. *L.Yuan, Z.Qiu*, "Mask-RCNN with spatial attention for pedestrian segmentation in cyber–physical systems", in Computer Communications, **vol. 180**, 2021, pp. 109-114

[27]. *P.Sedgwick*, "Spearman's rank correlation coefficient", in Bmj, 2014, pp. 349

[28]. *S.Patrick, S.Hannah, L.Bastian*, "Person attribute recognition with a jointly-trained holistic cnn model", in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 87-95

[29]. *J.J.Wu, H.Liu, J.G.Jiang, M.B.Qi, B.Ren, X.H.Li, Y.S.Wang*, "Person attribute recognition by sequence contextual relation learning", in IEEE Transactions on Circuits and Systems for Video Technology, **vol. 30**, no. 10, 2020, pp. 3398-3412

[30]. *D.Martinho-Corbishley, M.S.Nixon, J.N.Carter*, "Super-fine attributes with crowd prototyping", in IEEE transactions on pattern analysis and machine intelligence, **vol. 41**, no. 6, 2018, pp. 1486-1500