applied sciences

MDPI

Article

Pedestrian Attributes Recognition in Surveillance Scenarios Using Multi-Task Lightweight Convolutional Neural Network

Pu Yan [1,2], Li Zhuo [1,2,*], Jiafeng Li [1,2], Hui Zhang [1,2] and Jing Zhang [1,2]

[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; ppy@emails.bjut.edu.cn (P.Y.); lijiafenga@163.com (J.L.); huizhang@bjut.edu.cn (H.Z.); zhj@bjut.edu.cn (J.Z.)

[2] Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China

* Correspondence: zhuoli@bjut.edu.cn; Tel.: +86-136-5109-7164

Received: 23 August 2019; Accepted: 16 September 2019; Published: 7 October 2019

Abstract: Pedestrian attributes (such as gender, age, hairstyle, and clothing) can effectively represent the appearance of pedestrians. These are high-level semantic features that are robust to illumination, deformation, etc. Therefore, they can be widely used in person re-identification, video structuring analysis and other applications. In this paper, a pedestrian attributes recognition method for surveillance scenarios using a multi-task lightweight convolutional neural network is proposed. Firstly, the labels of the attributes for each pedestrian image are integrated into a label vector. Then, a multi-task lightweight Convolutional Neural Network (CNN) is designed, which consists of five convolutional layers, three pooling layers and two fully connected layers to extract the deep features of pedestrian images. Considering that the data distribution of the datasets is unbalanced, the loss function is improved based on the sigmoid cross-entropy, and the scale factor is added to balance the amount of various attributes data. Through training the network, the mapping relationship model between the deep features of pedestrian images and the integration label vector of their attributes is established, which can be used to predict each attribute of the pedestrian. The experiments were conducted on two public pedestrian attributes datasets in surveillance scenarios, namely PETA and RAP. The results show that, compared with the state-of-the-art pedestrian attributes recognition methods, the proposed method can achieve a superior accuracy by 91.88% on PETA and 87.44% on RAP respectively.

Keywords: pedestrian attributes recognition; lightweight convolutional neural network; multi-task; surveillance scenarios

1. Introduction

People often identify a person through discrete and precise attributes such as clothing style, gender, weight, and hairstyle. Attributes are types of high-level semantic features. Compared with low-level visual features, attributes are complex in terms of extraction and expression. It requires a lot of cost, labor and time to label the attributes, but they contain more abundant semantic information and have stronger robustness to illumination and angle changes. Therefore, the attributes can be exploited to finely characterize the appearance of pedestrians from multiple different aspects. Therefore, they have important value in the applications of person re-identification, video structuring analysis, etc. Pedestrian attributes recognition has also gradually become an important research direction in the field of machine vision.

Due to the uncontrollability of the application scenario, pedestrian attributes recognition faces many great challenges, including:

(1) The pedestrian images in the acquisition process are affected by many complex factors, for example, illumination, low image quality, multi-view, blur, and occlusion, which has a serious impact on pedestrian attributes recognition.
(2) Pedestrians' belongings, such as backpacks and luggage, also bring certain difficulties to pedestrian attributes recognition.
(3) Each pedestrian possesses different attributes, therefore, the data distribution of each attribute in the datasets is not balanced, such as wearing a hat, black clothing and purple clothing [1].

In essence, pedestrian attributes recognition is a multi-label classification problem, because pedestrian attributes are related to each other and are not completely exclusive. Many studies have been carried out on the problem of multi-label recognition of pedestrian attributes. On the whole, pedestrian attributes recognition generally adopts the framework of "feature extraction + classifier". According to the features extracted for attributes recognition, the development of pedestrian attributes recognition can be divided into two stages:

The first stage: Before 2012, handcrafted features were used to train the classifiers, such as Support Vector Machine (SVM), to obtain the pedestrian attributes recognition model. The handcrafted features used included global features and local features. The commonly used global features included low-level visual features such as color and texture, which are easily affected by some factors such as scene environment, occlusion, and illumination. The local features focus on the detailed information of the image and have more advantages in the representation of the pedestrian attributes than the global features.

Layne et al. [2] labeled 21 kinds of pedestrian attributes, including clothing style, gender, and hairstyle, and respectively trained the classifier for each attribute. In the training process, pedestrian images from different angles of the camera were selected to improve the robustness to angle changes.

Zhu et al. [3] established the APiS database and labeled 13 kinds of attributes, mainly for pedestrian attributes recognition in complex scenes. The AdaBoost classifier and K-NN (Nearest Neighbors) classifier were respectively adopted to perform the recognition of binary attributes and multiple attributes.

Schumann et al. [4] constructed a dataset for pedestrian detection, tracking, and re-identification, and also labeled 14 kinds of pedestrian attributes, including clothing, gender, and height. They then calculated the reliable recognition score of each attribute for person re-identification.

In conclusion, the pedestrian attributes recognition using handcrafted features has the following deficiencies:

(1) The handcrafted features are less robust to variations in environment, illumination, camera angle, etc.
(2) A classifier needs to be designed for the recognition of each attribute, so the complexity of implementation is very high.
(3) The intrinsic relationship among attributes is often ignored and not fully utilized.

The second stage: After 2012, deep learning has made important breakthroughs in many computer vision tasks such as target detection and tracking, image segmentation, and image classification. Convolutional neural networks (CNN) and other deep neural networks have been widely used to extract the deep features of pedestrians. With the assistance of the powerful and efficient representation of deep features, the recognition accuracy of pedestrian attributes can be improved significantly. A review of the deep learning-based approaches shows that not only the has the design of the network structure been studied, but the problem has also been studied from different aspects. Hence, there are many different research ideas, including global-based, local parts-based, sequential prediction-based, visual attention-based, curriculum learning-based, new designed loss function-based, and graphic model-based ideas. [1]. Below we list a few representative works.

For global-based methods, ACN [5] adopts a joint training CNN model and proposes multi-branch classification layers for each attribute. The model uses only the dependencies among the attributes, but not pedestrian gestures, context, etc.

Part-based methods often jointly utilize local and global information. For example, Wenhua Fang et al. [6] proposed a multi-task CNN method. First, according to the spatial location and semantic relationship of the attributes, the attributes are grouped into local and global attributes. Then, the two groups of attributes are classified using different CNN models in a multi-task manner.

For sequential prediction-based models, Jingya Wang et al. [7] proposed a joint recurrent learning (JRL) model, mining attribute context information and the relationship among attributes to improve the recognition accuracy. In this work, a new Recurrent Neural Network (RNN) coding–decoding network was specifically designed. The context information among pedestrians and the internal attributes of pedestrians were modeled together to learn an integrated network for pedestrian attributes recognition.

For loss function-based models, WPAL [8] first determines the location of different attributes by the weakly supervised method, obtains the attribute detection result, and then predicts the attributes in the detected results.

For curriculum learning-based methods, Sarafianos et al. [9] proposed the idea of combining the advantages of both multi-task and curriculum learning, introducing curriculum learning into the person attribute recognition task.

Yutian Lin et al. [10] designed an attribute person recognition (APR) network, which combines the pedestrian ID with labeled pedestrian attributes information, such as gender, hair, and clothing, to train the recognition model. Finally, the trained network model can be used to predict both pedestrian ID information and the pedestrian attributes. This method is also called a join verification network [11], which can improve the generalization ability of the network by combining ID loss and attribute loss.

In contrast to the previous attention-based methods, Tan et al. [12] incorporated parsing attention, label attention and spatial attention into a unified network for pedestrian attributes analysis. This method used different attention mechanisms to extract discriminative features, which are correlated and complementary, achieving more reliable attribute recognition.

In summary, compared to the methods using handcrafted features, the pedestrian attributes recognition methods using deep features have the following advantages:

(1) Deep neural networks (DNN) simulate the cognitive mechanism of the brain. Through training, the features from the lower layer to the upper layer of DNN are automatically extracted step by step from the big data. Finally, highly efficient deep feature representation can be obtained. The deep features have strong distinctive capability and are robust to various environmental and illumination changes.
(2) Using the multi-label learning method, the relationship among various attributes can be explored deeply, so that a much better recognition performance can be achieved.

In this paper, a pedestrian attributes recognition method is proposed by designing a multi-task lightweight deep convolutional neural network, which integrates the pedestrian attributes labels together, to accurately recognize pedestrian attributes in a unified framework. The network is composed of five convolutional layers, three pooling layers and two fully connected layers. The loss function is improved based on sigmoid cross-entropy, and the scale factor is added to balance the amount of various attributes data, thereby improving the recognition accuracy. The recognition accuracy on the two pedestrian attributes datasets, namely PETA and RAP, achieved 91.88% and 87.44%, respectively. Compared with the existing pedestrian attributes recognition methods, the proposed one can achieve state-of-the-art recognition accuracy.

2. The Proposed Pedestrian Attributes Recognition Method

Deep learning can obtain highly efficient feature representation through automatically learning from massive data. With its powerful feature representation and context information extraction, it has

achieved far more performance than traditional methods in the fields of image classification, speech recognition, and natural language processing.

This paper presents a pedestrian attributes recognition method by using a lightweight CNN, whose framework is shown in Figure 1. The network can be divided into three convolutional blocks, which are five convolutional layers, three pooling layers, and two fully connected layers. The network uses 3×3 convolution kernels, not only to ensure the receptive field, but also to reduce the number of parameters in the convolution layer. Batch normalization (BN) and the dropout layer are used several times in the network to improve the training performance and generalization of the model.

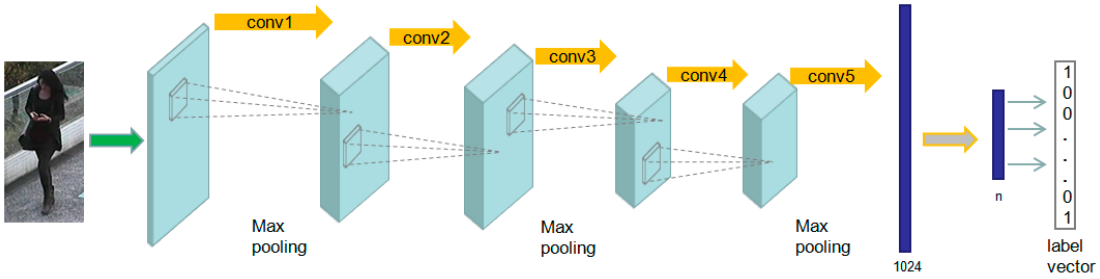


Figure 1. The framework of lightweight CNN used for pedestrian attributes recognition.

The BN layer can accelerate the training speed at the initial stage, guaranteeing that the training process can converge quickly, and effectively solves the problem of gradient divergence. Considering that the total number of images in the existing pedestrian attribute datasets is quite large, but the sample images of each attribute are not sufficient, a dropout layer is supplemented to the network, which stops the activation of a neuron with a certain probability during the forward propagation process. This strategy can enhance the generalization ability of the model and effectively suppress over-fitting. A BN layer follows each of the three convolutional blocks. The ratios of the three dropout layers are 25%, 25%, and 25%, respectively, and 50% after the first fully connected layer. The detailed definitions of the three convolutional blocks and the fully connected layers are shown in Table 1.

Table 1. The parameters of the layers in lightweight CNN.

Layer	Type	Kernel Size	Step Size	Kernels Number/Dropout Ratio	Output
Conv1	Convolutional layer	3×3	1	32	$96 \times 96 \times 32$
Pool1	Max pooling layer	3×3	3	-	$32 \times 32 \times 32$
Dropout1	Dropout layer	-	-	0.25	$32 \times 32 \times 32$
Conv2	Convolutional layer	3×3	1	64	$32 \times 32 \times 64$
Conv3	Convolutional layer	3×3	1	64	$32 \times 32 \times 64$
Pool2	Max pooling layer	2×2	2	-	$16 \times 16 \times 64$
Dropout2	Dropout layer	-	-	0.25	$16 \times 16 \times 64$
Conv4	Convolutional layer	3×3	1	128	$16 \times 16 \times 128$
Conv5	Convolutional layer	3×3	1	128	$16 \times 16 \times 128$
Pool3	Max pooling layer	2×2	2	-	$8 \times 8 \times 128$
Dropout3	Dropout layer	-	-	0.25	$8 \times 8 \times 128$
Flatten	Transition layer	-	-	-	$1 \times 1 \times 8192$
Fc1	Fully connected layer	1×1	1	1024	$1 \times 1 \times 1024$
Dropout4	Dropout layer	-	-	0.50	$1 \times 1 \times 1024$
Fc2	Fully connected layer	1×1	1	105	output

In the training phase, the pedestrian images and their corresponding attribute label integration vector are treated as a pairwise and input to the network. Through training, a mapping relationship model between the pedestrian image and the label integration vector is established. In the recognition phase, the input of the model is the pedestrian image, and the output is the label integration vector,

which respectively corresponds to the recognition result of each attribute of the pedestrian. Next, the integration manner of the pedestrian attributes labels will be described.

2.1. Integration of Pedestrian Attribute Labels

In this paper, the pedestrian's original attributes labels are integrated to form a vector. In this way, multiple attributes labels for each pedestrian image will be replaced by a vector.

Suppose there are N images to be converted, and L attributes corresponding to each image, including gender, age range, hair length, clothing color, clothing species, and so on.

Each sample is represented b $x_i, i \in [1, 2 \cdots, N]$, and the corresponding attribute vector is y_i. The attribute value corresponding to y_i is $y_{i,l}$, $y_{i,l} \in [0, 1], l \in [1, 2 \cdots, L]$. According to the original annotated label, if the pedestrian sample contains this attribute, then $y_{i,l} = 1$; otherwise $y_{i,l} = 0$.

In the PETA dataset, there are 61 binary attributes and four multi-class attributes. After combining the attributes of the pedestrian image, a label vector is generated, in which four multi-class attribute labels are also labeled in the form of binarization, and each multi-class attribute can be further divided into 11 binary attributes. Thus, each pedestrian image sample corresponds to a $61 + 11 \times 4 = 105$ dimensional label vector.

In the RAP dataset, there are 69 binary attributes and three multi-class attributes. The same processing method is adopted, and each image corresponds to a 92 dimensional label vector. In the process of integration, the position of each attribute in the label vector is fixed.

An example is shown to introduce how to integrate the attributes labels. A pedestrian sample in the PETA dataset is shown in Figure 2. The original labels are: upperBodyWhite, lowerBodyBlack, hairBlack, footwearWhite, lowerBodyCasual, lowerBodyTrousers, personalLess30, personalMale, upperBodyCasual, upperBodyLongSleeve, hairShort, footwearSneakers, carryingBackpack and accessoryMuffler. After integration, a 105-dimensional label vector is obtained, which is specifically shown as:

[0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0].

Figure 2. An example of a pedestrian image and labels in the PETA dataset.

2.2. Loss Function Design

In this paper, a sigmoid cross entropy-based loss function is adopted in the proposed model. Sigmoid is a kind of S-type function [13,14], and can represent the classification results in the form of output probability. Therefore, it can handle the multi-classification problem at the output of the neural network, and meet the requirements of pedestrian attributes recognition. As shown in Equation (1), where $p_{n,l}$ is the output probability of the lth attribute of the nth sample.

$$p_{n,l} = 1/(1 + \exp(-x_{n,l})) \tag{1}$$

Because the proposed model recognizes multiple attributes of the pedestrian sample simultaneously, it is essentially a multi-label recognition for pedestrians. Therefore, it is necessary to determine the relationship among attributes and consider the loss comprehensively. The overall sigmoid cross-entropy loss function is adopted in this paper, as shown in Equation (2).

$$Loss_{sigmoid} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L}\left(y_{n,l}\ln(p_{n,l}) + (1-y_{n,l})\ln(1-p_{n,l})\right) \quad (2)$$

The distribution of the samples is not balanced in the pedestrian datasets. In order to solve this problem, a positive sample scale index factor is introduced to the loss function to comprehensively determine the loss value of each attribute, and deal with the severe imbalance distribution of attributes. w_l represents the weight of the loss of the lth attribute, and the loss function with the positive sample scale factor added can be expressed by Equation (3).

$$Loss_{our} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L} w_l\left(y_{n,l}\ln(p_{n,l}) + (1-y_{n,l})\ln(1-p_{n,l})\right) \quad (3)$$

where p_l is the proportion of the positive sample of the lth attribute in the training set, and w_l can be calculated using the following equation:

$$w_l = \exp(-p_l/\sigma^2) \quad (4)$$

In this paper, the value of σ in Equation (4) is set to 1.

3. Experimental Results and Analysis

3.1. Datasets

At present, there are two commonly used pedestrian attributes datasets in surveillance scenarios, namely PETA [15] and RAP [16]. The pedestrian sample images in PETA are obtained by labeling several commonly used datasets. It contains 19,000 images of 8705 pedestrians. The resolution range is very large, from 17×39 to 169×365. The pedestrian attributes labels are given in a fixed form, containing 61 binary attributes and four multi-class attributes. There are 11 different color attributes in the multi-class, and some sample images of the PETA dataset are shown in Figure 3a.

The RAP dataset is currently the largest pedestrian attributes dataset, containing 41,585 images, all of which are taken indoors, with resolutions ranging from 36×92 to 344×554. In contrast to the PETA dataset, the pedestrian attributes labels in RAP contain 69 binary attributes and three multi-class attributes. In the multi-class labels, there are different numbers of color attributes. The dataset contains pedestrian images with different periods, different seasons and different perspectives and orientations. Some sample images of RAP are shown in Figure 3b.

While training the CNN model, the input image size is normalized to 96×96. The dataset is expanded by means of translation, folding, scaling, and random rotation of a certain angle to increase the number of samples of various attributes, and thus, to improve the recognition performance.

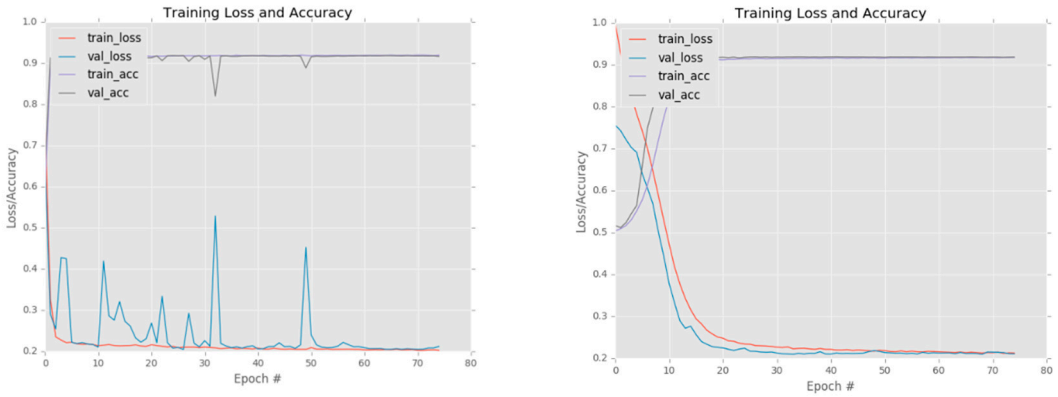
(**a**)

(**b**)

Figure 3. Examples of pedestrian images in the PETA and RAP datasets: (**a**) Sample images in the PETA dataset; (**b**) sample images in the RAP dataset.

3.2. CNN Parameter Settings

The CNN network is trained by using the random gradient descent method. The initial learning rate is 0.0001, the weight decay is set to 0.005, the batch size is 64, and 75 epochs are trained. The dataset is divided into a training set and test set according to a proportion of 80% and 20%, respectively. Through experiments, we found that the initial learning rate has a significant impact on the training process of the proposed model, which is manifested as the "gradient jitter", as shown in Figure 4a,b. It can be seen that when the initial learning rate is 0.001 in Figure 4a and 0.0001 in Figure 4b, there is a significant jitter.



(**a**) (**b**)

Figure 4. The impact of learning rate on performance during training with (**a**) a 0.001 initial learning rate, and (**b**) a 0.0001 initial learning rate.

At present, most of the literature adopts mean Accuracy (mA) as the evaluation metric of attributes recognition algorithms [1,5,6]. Therefore, we also adopted mA as the metric to measure the recognition performance in this paper. The calculation process of mA is as follows.

For each pedestrian attribute, the recognition accuracy of positive and negative samples is calculated separately, and then the average value is taken as the final recognition accuracy of the attributes. After that, the average value of the total pedestrian attributes recognition accuracy will be used as the final recognition rate, that is, mA is calculated as follows:

$$mA = \frac{1}{2L}\sum_{i=1}^{L}\left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}\right) \tag{5}$$

where L represents the number of attributes, P_i represents the number of positive samples of the ith attribute in test samples, TP_i represents the number of correctly predicted positive labels of the ith attribute in test samples, N_i represents the number of negative samples of the ith attribute in test samples, and TN_i represents the number of correctly predicted negative labels of the ith attribute in test samples.

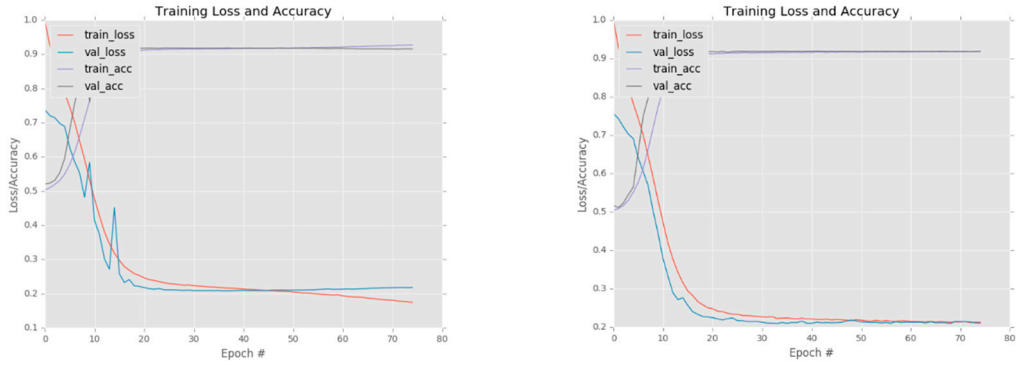
3.3. Impact of Data Enhancement on Recognition Accuracy

Although PETA is currently a large pedestrian attributes dataset containing 19,000 images, the sample data scale is not large enough for each attribute. Therefore, the dataset is expanded to 190,000 images by means of horizontal translation, vertical translation, horizontal folding, scaling transformation, and random rotation of a certain angle. Table 2 shows the effect of data enhancement on recognition accuracy.

Table 2. The impact of data enhancement on recognition accuracy.

Data Enhancement or Not	Loss	mA
no	0.2181	91.52%
yes	0.2167	91.88%

The experimental results show that mA can be improved by 0.36% through data enhancement. The results of the loss and accuracy in the training process are shown in Figure 5, where Figure 5a is without data enhancement, and Figure 5b with data enhancement, respectively. It can be seen that after the data enhancement, the jitter of the loss reduction in the whole training process becomes smaller and also tends to be earlier, and the accuracy of the training set or test set is relatively high. Therefore, it can be concluded that data enhancement has a good impact on recognition accuracy and training convergence.



(**a**) (**b**)

Figure 5. The loss and accuracy of the training process with data enhancement on the PETA dataset: (**a**) without data enhancement; and (**b**) with data enhancement.

3.4. Comparisons with State-of-the-Art Methods

In order to verify the performance of the proposed method on the PETA and RAP datasets, we compared it with the two traditional pedestrian attribute recognition representative methods; ikSVM [2,17] and ELF [18], and another four deep learning-based methods, including ACN [5], DeepMAR [19], WPAL [8] and multi-task CNN [6].

The traditional ikSVM method uses the handcrafted features and the SVM classifier framework to recognize the attributes, and a classifier is trained for each attribute. In order to achieve a better recognition performance, the number of positive samples or negative samples is intentionally balanced in an artificial manner.

ELF [18] (ensemble of localized features), a method of using local features, combines the features of eight color channels and luma channels to obtain feature representations and then uses AdaBoost as the classifier.

In deep learning-based methods, ACN [5] adopts a joint training CNN model that uses only the dependencies among attributes, but not pedestrian gestures, context, etc. Furthermore, a N/A label is added, for the first time in the labeling process, which denotes the uncertainty of certain attributes.

The DeepMAR [19] method has often been used as the benchmark. This paper proposes two models. One is for the recognition of a single attribute and the other for multiple attributes. A simple CNN model is proposed for feature extraction and joint training of multiple attributes. The recognition accuracy has been improved by using the correlation of attributes.

In contrast to previous methods of attributes recognition using the whole pedestrian image, WPAL [8] first determines the location of different attributes by the weakly supervised method, obtaining the attribute detection result, and then predicts the attributes in the detected results.

Among the existing deep learning-based pedestrian attributes recognition methods, the multi-task CNN [6] method can achieve state-of-the-art recognition performance. The basic idea is to classify pedestrian attributes in a multi-task manner.

Table 3 shows a comparison of the recognition accuracy of the proposed method and the other five methods on the PETA dataset.

Table 3. Comparison results using different methods on the PETA dataset.

Methods	mA
ikSVM	82.75%
ELF	75.21%
ACN	81.15%
DeepMAR	82.89%
WPAL	85.50%
multi-task CNN	88.20%
Ours	91.88%

As shown in Table 3, compared with the current deep learning-based pedestrian attributes recognition methods, the proposed method can obtain the highest recognition accuracy of 91.88%, which is 3.68% higher than that of multi-task CNN.

We also conducted experiments on the RAP dataset. The attribute labels were divided into two groups according to the behavior attribute and the item attribute, and a 92-dimensional label vector was generated. The comparison results of the recognition accuracy on the RAP dataset are shown in Table 4.

Table 4. Comparison results of different methods on the RAP dataset.

Methods	mA
ELF	69.94%
ACN	69.66%
DeepMAR	73.79%
WPAL	81.25%
multi-task CNN	83.25%
Ours	87.44%

As shown in Table 4, compared with the current deep learning-based pedestrian attributes recognition methods, the proposed method can also obtain the highest recognition accuracy of 87.44% on the RAP dataset, which is 4.21% higher than that of multi-task CNN.

The main reasons why the proposed method can achieve superior recognition accuracy are:

(1) The unique design of the BN layer and the dropout layer are adopted, which results in the improvement of recognition accuracy of the attributes.
(2) The integration of the attributes labels to form a label vector is equivalent to re-assigning a new label to each pedestrian image. By using CNN, the mapping relationship between the pedestrian image and the label vector can be established with high efficiency. The label vector combines all the attributes together, and thus, can fully exploit the intrinsic correlation among the attributes, improving the recognition accuracy.
(3) The imbalance problem of the attribute samples is solved by using the improved loss function.

3.5. Attributes Recognition Result

Using the proposed model, the pedestrian attributes can be recognized. The results are shown in Figure 6, where the prediction confidence of 10 attributes is the output. It can be seen that the accurate attribute recognition can be obtained on the pedestrian sample.

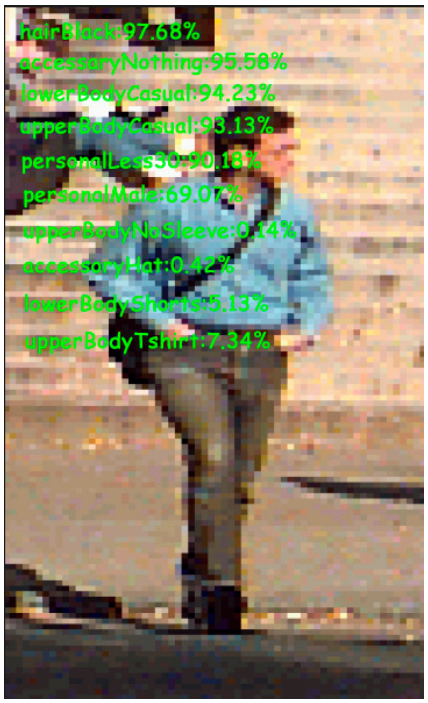


Figure 6. Recognition results of 10 attributes.

4. Conclusions

In this paper, a pedestrian attributes recognition method is proposed using a lightweight CNN. Considering the intrinsic relationship among the pedestrian attributes, the recognition task of pedestrian attributes can be completed in a unified framework. On the public pedestrian attribute dataset PETA, the mean recognition accuracy of the proposed method is as high as 91.88%, and is 87.44% on the RAP dataset. Compared with the existing deep learning-based pedestrian attributes recognition methods, the proposed method obtains a superior performance. In future work, we will also improve the recognition accuracy of pedestrian attributes by adjusting the hyper parameters and grouping the attributes, and we will further apply the pedestrian attributes to person re-identification.

Author Contributions: Conceptualization, P.Y. and L.Z.; methodology, L.Z.; software, P.Y.; writing—original draft preparation, P.Y.; writing—review and editing, P.Y. and L.Z. and J.L.; supervision, H.Z. and J.Z.

Funding: This work in this paper was supported by the Beijing Municipal Natural Science Foundation Cooperation Beijing Education Committee (No. KZ 201810005002, No. KZ 201910005007), National Natural Science Foundation of China (No. 61531006, No. 61971016, No. 61602018, and No. 61701011).

Acknowledgments: The authors would like to thank all the colleagues who contributed to this research work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, X.; Zheng, S.; Yang, R.; Luo, B.; Tang, J. Pedestrian Attribute Recognition: A Survey. *arXiv* **2019**, arXiv:1901.07474.
2. Layne, R.; Hospedales, T.M.; Gong, S. *Attributes-Based Re-Identification*; Person Re-Identification; Springer: London, UK, 2014; pp. 93–117.
3. Zhu, J.; Liao, S.; Lei, Z.; Yi, D.; Li, S.Z. Pedestrian Attribute Classification in Surveillance: Database and Evaluation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, NSW, Australia, 2–8 December 2013; pp. 331–338.
4. Schumann, A.; Monari, E. A soft-biometrics dataset for person tracking and re-identification. In Proceedings of the 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, Korea, 26–29 August 2014; pp. 193–198.
5. Sudowe, P.; Spitzer, H.; Leibe, B. Person attribute recognition with a jointly-Trained holistic CNN model. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 329–337.
6. Fang, W.; Chen, J.; Hu, R. Pedestrian attributes recognition in surveillance scenarios with hierarchical multi-task CNN models. *China Commun.* **2018**, *15*, 208–219.
7. Wang, J.; Zhu, X.; Gong, S.; Li, W. Attribute Recognition by Joint Recurrent Learning of Context and Correlation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 531–540.
8. Zhou, Y.; Yu, K.; Leng, B.; Zhang, Z.; Li, D.; Huang, K.; Feng, B.; Yao, C. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv* **2017**, arXiv:1611.05603.
9. Sarafianos, N.; Giannakopoulos, T.; Nikou, C.; Kakadiaris, I.A. Curriculum learning for multi-task classification of visual attributes. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop, Venice, Italy, 22–29 October 2017; pp. 2608–2615.
10. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Yang, Y. Improving person re-identification by attribute and identity learning. *arXiv* **2017**, arXiv:1703.07220, 2017. [CrossRef]
11. Zheng, Z.D.; Zheng, L.; Yang, Y. A discriminatively learned CNN embedding for person re-identification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 13.
12. Tan, Z.; Yang, Y.; Wan, J.; Hang, H.; Guo, G.; Li, S.Z. Attention-Based Pedestrian Attribute Analysis. *IEEE Trans. Image Process.* **2019**, *12*, 6126–6140. [CrossRef] [PubMed]
13. Hassanzadeh, H.R.; Rouhani, M. A Multi-Objective Gravitational Search Algorithm. In Proceedings of the 2012 2nd International Conference on Computational Intelligence, Communication Systems and Networks, CICSyN, Liverpool, UK, 28–30 July 2010; pp. 7–12.

14. Kendall, A.; Gal, Y.; Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.
15. Deng, Y.; Luo, P.; Loy, C.C.; Tang, X. Pedestrian attribute recognition at far distance. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 789–792.
16. Li, D.; Zhang, Z.; Chen, X.; Ling, H.; Huang, K. Arichly annotated dataset for pedestrian attribute recognition. Computer Vision and Pattern Recognition. *arXiv* **2016**, arXiv:1603.07054.
17. Zhu, J.; Liao, S.; Lei, Z.; Li, S.Z. Multi-label convolutional neural network based pedestrian attribute classification. *Image Vision Comput.* **2016**, *58*, 224–229. [CrossRef]
18. Gray, D.; Tao, H. Viewpoint invariant pedestrian recognition with an ensemble of localized Features. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision (ECCV), Marseille, France, 12–18 October 2008*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 262–275.
19. Li, D.; Chen, X.; Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 111–115.