Research on Pedestrian Attribute Recognition Based on Semantic Segmentation in Natural Scene

Xin Feng[1], Yangyang Li[2], Haomin Du[1], and Hongbo Wang[1(✉)]

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
xfeng@bupt.com, {duhaomin,hbwang}@bupt.edu.cn
[2] National Engineering Laboratory for Public Safety Risk Perception and Control, CAEIT, Beijing 100041, China
liyangyang@cetc.com.cn

Abstract. Smart city is a new term given to society by technology, and cameras are important infrastructure for building a smart city. How to use camera information efficiently and effectively plays an important role in people's daily life and maintain social order. Pedestrian information accounts for a large proportion of camera information, so we hope to make good use of pedestrian information. Previous works use traditional machine learning methods and neural network to identify pedestrian attributes, mainly judge the existence of pedestrian attributes in natural scenes. However, it's not enough to judge whether an attribute exist or not, getting the position of an attribute often gives you more information. In this paper, we propose to use semantic segmentation to obtain the position information of pedestrian attributes. We first propose pedestrian attribute semantic dataset in natural scene called PASD (Pedestrian attribute semantic dataset), which select 27 visualized pedestrian attributes. Deeplabv3+ is used to perform experiments on PASD, which obtain the mIoU (mean intersection over union) baseline of 27 pedestrian attributes. For getting useful conclusion, we conduct data analysis about mIoU from three aspects: attribute distribution, accuracy and resolution.

Keywords: Pedestrian attribute · Location information · Semantic segmentation

1 Introduction

With the development of artificial intelligence like image classification [18] and internet of things technology, Smart cities have been well developed. Urban cameras have become more and more popular, and the analysis of camera information is significant to public safety. Pedestrian attribute is an important information type in camera information and can be applied to pedestrian search [1], pedestrian recognition [2, 3], and pedestrian re-recognition [4].

Hence, pedestrian attribute recognition has important practical significance. Previous work used traditional machine learning methods and neural network to identify

X. Sun et al. (Eds.): ICAIS 2019, LNCS 11635, pp. 498–509, 2019.
https://doi.org/10.1007/978-3-030-24268-8_46

pedestrian attributes, mainly judge the existence of pedestrian attributes in natural scenes.

However, it's not enough to judge whether an attribute exist or not, getting the position of an attribute often gives you more information. For example, the position information of pedestrian helps to improve the ability of pedestrian re-recognition. As we know, human view the position of attributes as supplementary information to judge whether two pedestrians are same person, as shown in Fig. 1. We can conduct pedestrian re-recognition by using hair, tops and pant in different position. So, when machine get this information, it will improve the pedestrian re-recognize task. On the other hand, if obtaining the location of attributes, we can do further research for some attributes with known location information, like analyzing the brand of bag, the size of shoes and so on.

Fig. 1. Re-identification of attributes by attribute location information (Color figure online)

There is a challenge in getting pedestrian attribute position: The diversity of pedestrian postures leads to the diversity of pedestrian attributes. For example, the regional information of pedestrian clothes may change with the pedestrian's swing. Therefore, if you obtain the position information of the pedestrian attribute, you need to consider how to express the pedestrian attribute location information. First, we considered the general target detection algorithm, mainly researching an improved method based Faster R-CNN [6] and YOLO (You only look once). During the investigation, it was found that the general target detection algorithm detects the object by marking the object in the form of a rectangular frame, which will cause a lot of repetitive areas and can't represent the location information of attributes accurately. To solve upon problem, we turn our attention to semantic segmentation.

Semantic segmentation is a pixel-level image classification method. By classifying each pixel of a picture, the position information of each category in the picture can be obtained. Taking Fig. 1 as an example, the pixels belonging to the hair should be classified into hair and signed with blue, the pixels belonging to the pants will be signed with bright green to represent pants. Different classifications are signed with different colors. At the moment, we can see that no matter how the pedestrian attribute changes, semantic segmentation can clearly distinguish the outline of the attribute.

This paper proposes to use the method of semantic segmentation to obtain the pedestrian attributes and their position information in natural scenes. We will verify its feasibility, explore the problems and valuable rules that need to be paid attention to when use this method. In this paper, mIoU (mean intersection over union) is used to measure the performance of semantic segmentation model.

The main contributes of this paper is: (1) we proposed the first pedestrian attribute semantic dataset in natural scene called PASD (Pedestrian Attribute Semantic Dataset). (2) We use semantic segmentation framework deeplabv3+ to conduct experiment and get the mIoU reference value of pedestrian attributes, which can provide reference for the work afterwards. (3) In this paper, mIoU is analyzed from three aspects: accuracy, percentage of PASD pedestrian attribute categories, and resolution of the image. The factors influencing mIoU of pedestrian attributes are summarized.

2 Related Work

2.1 The Research of Pedestrian Attribute Recognition

With the development of neural network, the neural network method achieves higher accuracy in pedestrian attribute recognition than traditional machine learning. There are three research directions based on neural network research: the first research direction is based on traditional neural networks. Mainly through the neural network itself to identify attribute features. DeepSAR and DeepMAR [7] can make better use of the correlation between various attributes; the second research direction is to transform attribute gender into serialization model. Neural PAR [8] transforms attribute recognition problem into attribute generation problem, and uses LSTM for attribute recognition. The model of JRL [15] based on the RNN network model explicitly explores a sequential prediction constraint; the third research direction is to apply the attention mechanism to attribute recognition. Sarafianos [16] and HydraPlus-Net [17] both use the multi-level features to better identify pedestrian attributes.

2.2 Semantic Segmentation

Semantic segmentation is the pixel-level classification of images. By classifying each pixel of an image, the position information of each category in the picture can be obtained. With the success of convolution network, the convolution network is quickly applied to semantic segmentation.

Recently, the encoder-decoder network structure is very common in semantic segmentation. The encoder-decoder consists of two parts: (1) the spatial dimension of the feature map will gradually decrease in the encoder part, while the longer range information is easier to capture. (2) The decoder part will gradually restore the object details and spatial dimensions. For example, SegNet [13] uses encoder's pooling indices to learn additional convolutional layers. U-Net [14] added a skip connection between the encoder and the corresponding layer of decoder. Deeplab [9, 10] use Atrous convolution, proposing Atrous Spatial Pyramid Pooling and combining with the latest network structure, which further improved the accuracy of semantic segmentation.

In the use of semantic segmentation for attribute recognition, Kalayeh et al. [11] proposed an encoder-decoder based SSP and SSG structure to achieve recognition of face attributes. This paper focuses on the semantic segmentation of pedestrian attributes in natural scenes. Compared with face attribute segmentation, pedestrian attribute categories have more complexity and therefore have greater challenges.

3 Method

3.1 The Choice of Semantic Segmentation Framework

When judge to select a semantic segmentation model, we mainly consider two factors: accuracy and efficiency. In terms of accuracy, this paper uses the mIoU of each model in the PASCAL VOC2012 test set as the benchmark. Deeplabv3+ [10] get the best mIoU with 87.8.

In terms of model efficiency, many semantic segmentation models do not study the efficiency problem. The author mainly considers the use of deep separable convolution. The depth separable convolution can reduce the number of convolution parameters while maintaining the effect of extracting features. All convolutions in the Xception model used by deeplabv3+ use deep separable convolution, which greatly reduces the size of the model and greatly increases the computational efficiency of the model.

Based on the above factors, we use deeplabv3+ as a basic model to analyze pedestrian attributes. This model combines the advantages of deeplabv3 and encoder-decoder.

3.2 The Introduction of Deeplab Structure

Deeplabv3 contains key structures with Atrous convolution and ASPP (Atrous Spatial Pyramid Pooling). Deeplabv3+ adds Xception [12] to the basic model, which greatly improves the training efficiency.

Atrous Convolution. Atrous convolution is an improvement to ordinary convolution. For an input of 9 * 9, if use ordinary convolution, you will lose a lot of spatial information. Atrous convolution preserves the spatial information of the image without adding parameters by inserting zero into the convolution kernel. Atrous convolution can predict any precision in the final convolution layer, which can more easily control the size of input and output in the convolution network then ordinary convolution.

Consider a two-dimensional case, assuming that the input size is m * m, the size of convolution kernel is k, and the step size is s. A new parameter r has been introduced in Atrous convolutions, assume that the output size is n * n. There are following formulas:

$$n = (m + (k - 1)(r - 1) - k)/(2 * s) \tag{1}$$

ASPP. In semantic segmentation, the size of each category is different. In order to solve this problem, deeplab draws on the success of spatial pyramiding in the visual field, using Atrous Spatial Pyramid Pooling. The special of ASPP is to use multiple Atrous convolutions with different rates in parallel and concat the results of each branch to get the final result. Atrous Spatial Pyramid Pooling is shown in Fig. 2.

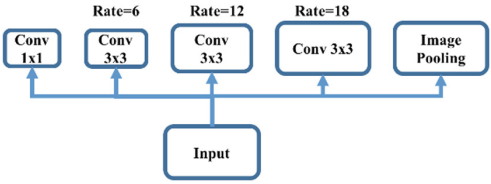

Fig. 2. The structure of Atrous Spatial Pyramid Pooling

In addition, deeplabv3+ adds an image-level feature. Image pooling construct of three part: First, a global average pooling operation is performed on the input of the final feature matrix, and then the result of pooling passes a 1 * 1 convolution with 256 channels followed by a batch normalization. Finally, the feature is restored to the desired size by bilinear upsampling.

Depthwise Separable Convolution. Depthwise separable convolution separate original convolution into depthwise convolution and pointwise convolution. The structure of depthwise separable convolution is shown as Fig. 3. Deeplabv3+ also uses depthwise separable convolution in Atrous Spatial Pyramid Pooling, which gain better computation efficiency with guaranteed accuracy.

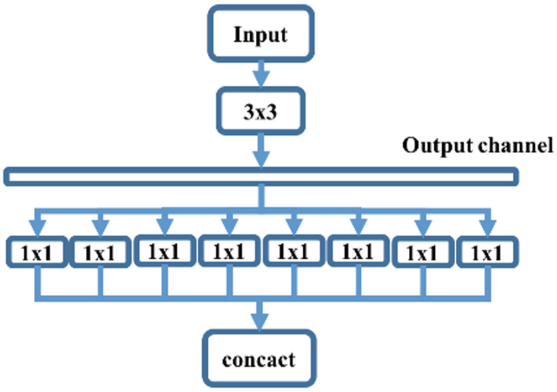

Fig. 3. Depthwise separable convolution

Xecption. The original Xception uses 36 convolution layers. This network is divided into three parts: entry flow, middle flow and exit flow. Deeplabv3+ use an improved version of Xception and made the following three changes: (1) Deeper Xception same as in [12] except do not modify the entry flow network structure for fast computation and memory efficiency. (2) All max pooling operations are replaced by depthwise separable convolution with striding, which enables network to extract feature maps at an arbitrary resolution. (3) Adding ReLU and batch normalization after each 3 * 3 depthwise convolution.

Encoder-Decoder. Deeplabv3+ uses deeplabv3 as encoder. Deeplabv3 uses Atrous convolution to extract feature maps at an arbitrary resolution. There introduces a new parameter output stride to represent the ratio of the input picture size to the output size. For example, the final feature map is usually 1/32 of the input and the value of output stride is 32. In semantic segmentation, output stride is often set as 16 or 8. Deeplabv3+ uses an improved Atrous Spatial Pyramid Pooling, which add an image level feature (implemented by an avgpool) and concat with Atrous convolution with different rates. This concentrated value is as the output of encoder.

In the part of decoder, deeplabv3+ proposes a decoder structure suitable for semantic segmentation. First, the output of encoder is upsampled with factor of 4, then it is merged with the low-level features in the encoder containing the same space size and use 1×1 convolution to reduce the number of channel which make training easily as the low-level feature of encoder always contains a lot of channels.

Loss function. Deeplabv3+ use cross-entropy softmax as loss function. Deeplabv3+ model used by our experiment is shown as Fig. 4.

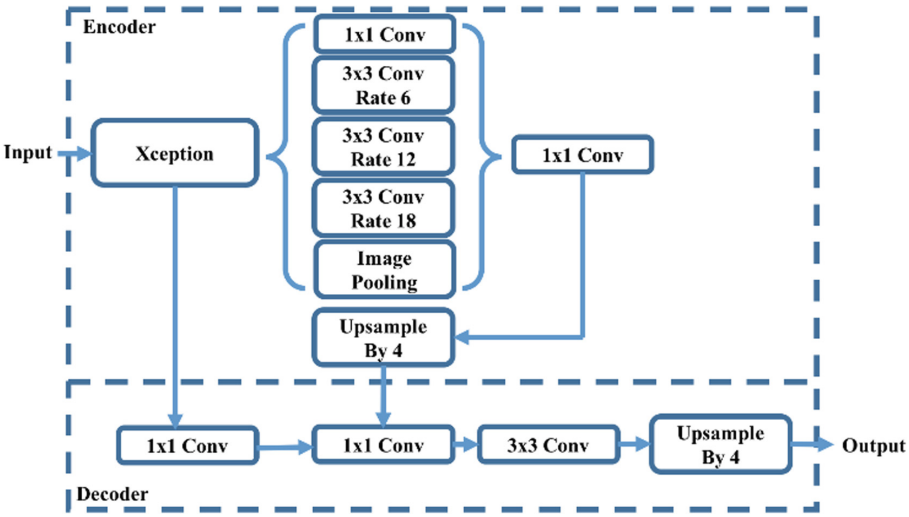

Fig. 4. The structure of deeplabv3+

4 Dataset

So far, there is no available dataset for semantic annotation of pedestrian attributes in natural scenes for experimental. Therefore, we created the first pedestrian attribute dataset with semantic annotations in natural scenes called PASD (Pedestrian attribute semantic dataset). First of all, we select 1461 pedestrian images from the PETA [5] for semantic annotation. PETA is a dataset in the field of pedestrian recognition in natural scenes accepted by academy. It contains 19000 images with 35 pedestrian attribute categories. In selecting image, we consider the following factors: (1) Image resolution,

we select semantically segmented images based on human recognition and ignore images that are not recognized by human. (2) The balance of categories. For some categories have few number (such as stripe, grid), we try to maintain the balance of this categories.

In order to analyze the effect of resolution on pedestrian attribute recognition using semantic segmentation. In addition to selecting images from PETA, we also selected 250 pedestrian images taken by city surveillance cameras in real life scenarios. The resolution of these images are higher than those in PETA.

For classification selection, we first refer to the pedestrian attribute selection method of the PETA. PETA divides the pedestrian attributes into 35 categories which are based on attributes that is selected by anthropologists to represent human characteristics, and the attribute categories of the dataset itself. We consider the distribution of attributes in the training set, and certain attributes (such as age, gender, etc.) cannot be separated and visualized by semantic segmentation. Finally, the author selects 27 pedestrian attributes. The attribute values are shown in Table 1.

Table 1. The pedestrian attributes we finally selected

Sunglasses	Jeans	Sneaker	Leather shoes
Shorts	Stripes	Logo	Trousers
ShortSleeve	Tshirt	Skirt	Backpack
Hat	Casual upper	Hair	Sandals
Plaid	Suitcase	Bag	Muffler
Formal upper	Long hair	Shoes	Jacket
Plastic bag	CarryingOther	V-Neck	

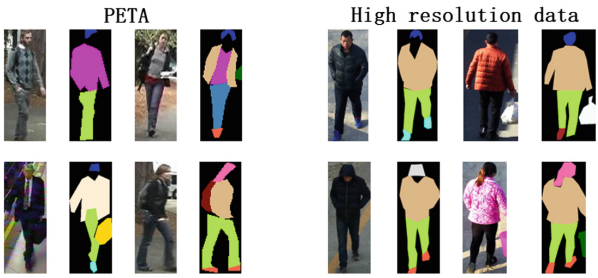

Fig. 5. Some images in our dataset, which is consist of parts of PETA and some high resolution images.

PASD consist of 1,711 semantic segmentation annotations, of which there are 1461 PETA datasets and 250 datasets of our own. Some images in PASD are shown in Fig. 5.

5 Experiment

5.1 Experiment Setup

For the input data, no additional operation is performed on the picture except resize the picture to 513. The initial weight, we set the basic learning rate to 0.001, and the learning rate is reduced using poly strategy. We use momentum and set the number of iterations to 900000. The batch norm layer is not trained. The value of output stride that we train and test is set to 8. The corresponding ASPP parameters are set to 6, 12 and 18 respectively. To prevent overfitting, the dropout's keep value is set to 0.3. At the same time, we adopt the early stop strategy. GPU is 1080ti.

5.2 Experiment Results

We randomly selected 1273 images as training data. The test data selected 288 images containing 188 low resolution images and 100 high resolution images. The mIoU of 288 test set images is shown in Table 2.

Table 2. mIoU of 288 test set images

Attribute	mIoU	Attribute	mIoU
Sunglasses	0	Jeans	0.378738642
Sneaker	0.354543	Leather shoes	0.184471875
Formal upper	0.603124	Shorts	0.346722752
Stripes	0.519705	Logo	0
Trousers	0.832277	Long hair	0.6549505
ShortSleeve	0.473253	Tshirt	0.26975289
Skirt	0.563472	Backpack	0.233864143
Shoes	0.390536	Hat	0.229438677
Casual upper	0.444053	Hair	0.66051203
Sandals	0.196731	Jacket	0.731595635
Plaid	0.458361	Suitcase	0.731545508
Bag	0.416399	Muffler	0.159364507
Plastic bag	0.350488	CarryingOther	0.142766684
V-Neck	0.022614		
Average	0.31996		

6 Data Analysis

It can be seen that the overall mIoU is still relatively low, and there is a severe overfitting in the training phase. We analyzed the mIoU value from the following three aspects.

6.1 The Relationship Between mIoU and Each Category Ratios

Firstly, we give a line chart of the relationship between mIoU and category ratios, as shown in Fig. 6.

It can be seen that when the number of samples is small, the change trend of mIoU and category ratios is same basically. After analyzing the pictures and the Table 2, the following factors affect the category mIoU: (1) A relatively high mIoU can be obtained if the category profile is fixed or has unique visualization features. Such as Suitcase, strips, Jacket. Conversely, a relatively low mIoU is obtained for categories with more category diversity. Such as shoes, which has variability and uncertainty in the natural scene, that is, the human eye can't tell which kind of shoes it is. (2) Image resolution, such as sunglasses, most of the sunglasses in the data set are difficult to identify clearly. (3) The proportion of categories still plays a key role. We have calculated the correlation coefficient between the proportion of mIoU and category ratios is 0.492077145, so the proportion of mIoU and category ratios is still positively correlated.

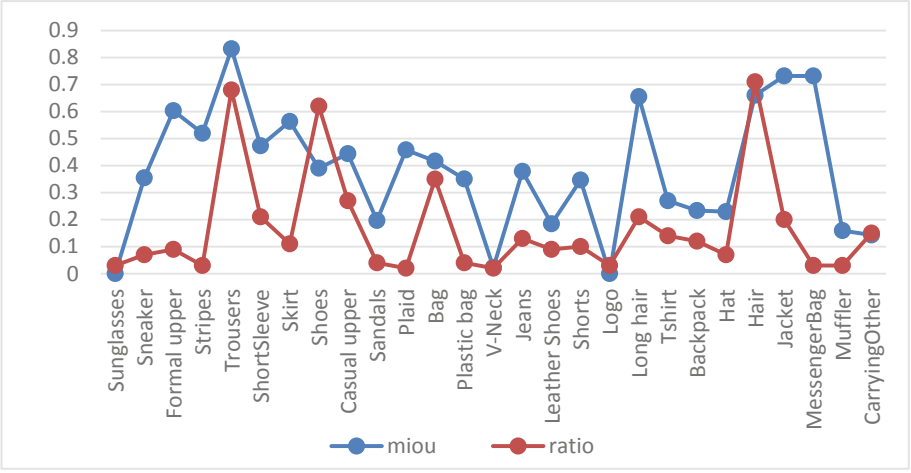

Fig. 6. The relationship between mIoU and each category ratios

6.2 The Relationship Between mIoU and Accuracy

In addition to mIoU, we also output 27 types of prediction accuracy. The method for calculating the accuracy here is that if the number of accumulated pixels of a certain category in the image is greater than 10, this image is considered to contain the category. The calculated accuracy of the 288 test set images is shown in Table 3.

We see that the accuracy rate of only one of the Trousers exceeded the result in [8]. In order to analyze the reasons, we plotted the relationship between mIoU and accuracy, as shown in the Fig. 7.

Table 3. The accuracy of pedestrian attributes

Attribute	mIoU	Attribute	mIoU
Sunglasses	0	Jeans	0.457143
Sneaker	0.770992	Leather Shoes	0.346154
Formal upper	0.692308	Shorts	0.636364
Stripes	0.535714	Logo	0
Trousers	0.985294	Long hair	0.803571
ShortSleeve	0.728571	Tshirt	0.512821
Skirt	0.833333	Backpack	0.527778
Shoes	0.666667	Hat	0.56
Casual upper	0.692308	Hair	0.936893
Sandals	0.538462	Jacket	0.875
Plaid	0.384615	Suitcase	0.461538
Bag	0.747573	Muffler	0.4
Plastic bag	0.636364	CarryingOther	0.537037
V-Neck	0.2		

We calculated the correlation coefficient between mIoU and accuracy to be 0.78534118, which obtain a positive correlation between mIoU and accuracy. That is, mIoU directly affects accuracy, so you can improve accuracy by raising mIoU.

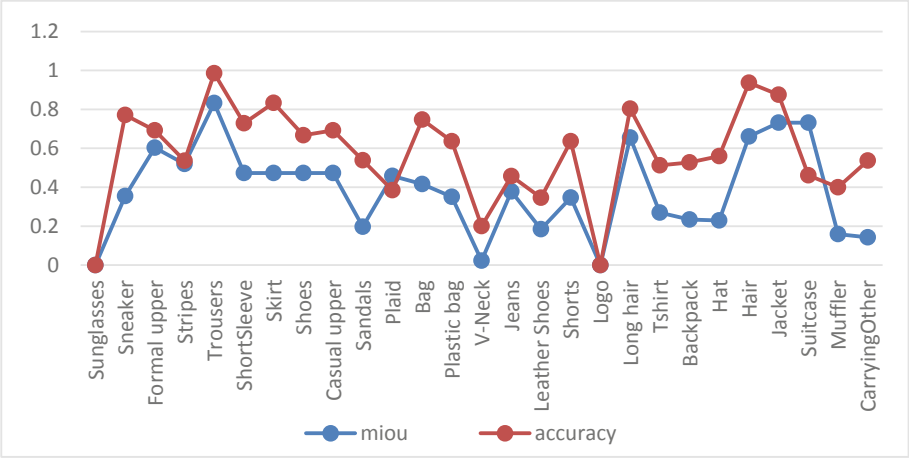

Fig. 7. The relationship between mIoU and accuracy

6.3 The Relationship Between mIoU and Resolution

Finally, in order to test the effect of different resolution on mIoU. We selected 100 high resolution images from the surveillance screen for testing. Note that the selected high resolution dataset contains only 12 categories. We draw low resolution images and high resolution images mIoU relationship table, as shown in the Fig. 8.

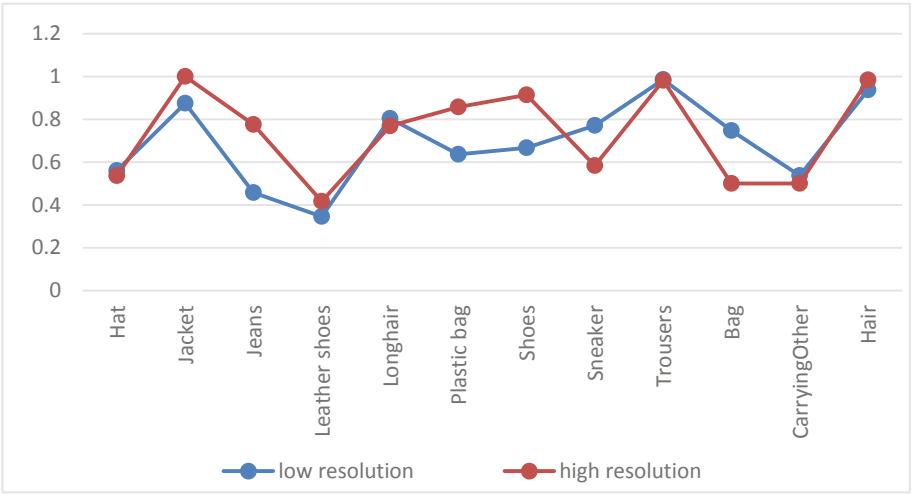

Fig. 8. The relationship between mIoU and resolution

As you can see, the high resolution mIoU is generally higher than the low resolution mIoU. For Longhair and Sneaker, there are special circumstances. We analyzed that class diversity leads to poorly recognized issues that make Long hair recognized as a hair and Sneaker recognized as shoes.

7 Conclusion

This paper focus on obtaining the location information of pedestrian attributes by semantic segmentation. We prove that semantic segmentation is feasible for obtaining pedestrian attributes. We propose the first pedestrian attribute semantic dataset PASD and use deeplabv3+ to get mIoU baseline with low resolution images and high resolution images.

In addition, we analyzed the experimental data and draw some conclusion. Firstly, mIoU is proportional to the number of samples when there are few data samples. In particular, attributes with variability are not well divided. The Second is the problem of data resolution: as the society develops, the resolution of the camera will continue to increase, and the low resolution images in the PETA are not applicable in reality. Therefore, some images of relatively high resolution should be obtained. It was verified in experiments that semantic segmentation yields better results on high resolution images.

Acknowledgement. This work was supported by CETC Joint Research Program under Grant 6141B08020101, 6141B0801010a, and the National Natural Science Foundation of China under Grant 61002011.

References

1. Sami, J.E., Nixon, M.: Analysing soft clothing biometrics for retrieval. In: International Workshop on Biometric Authentication, pp. 234–245 (2014)
2. Martinson, E., Lawson, W., Trafton, J.G.: Identifying people with soft-biometrics at fleet week. In: IEEE International Conference on Human-Robot Interaction, pp. 49–56 (2013)
3. Reid, D., Nixon, M., Stevenage, S.: Soft biometrics: human identification using comparative descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1216–1228 (2014)
4. Li, A., Liu, L., Wang, K., Liu, S., Yan, S.: Clothing attributes assisted person re-identification. TCSVT **25**, 869–878 (2015)
5. Deng, Y., Luo, P., Loy, C.: Learning to recognize pedestrian attribute. arXiv preprint arXiv:1501.00901 (2015)
6. Meng, R., Rice, S.G., Wang, J., Sun, X.: A fusion steganographic algorithm based on faster R-CNN. CMC: Comput. Mater. Continua **55**(1), 001–016 (2018)
7. Li, D., Chen, X., Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: IEEE Asian Conference on Pattern Recognition, pp. 111–115 (2015)
8. Ji, Z., Zheng, W., Pang, Y., Deep pedestrian attribute recognition based on LSTM. In: IEEE International Conference on Image Processing (2017)
9. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous convolution for semantic image segmentation. In: CVPR (2017)
10. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: CVPR (2017)
11. Kalayeh, M.M., Gong, B., Shah, M.: Improving facial attribute prediction using semantic segmentation. In: CVPR (2017)
12. Qi, H., et al.: Deformable convolutional networks – COCO detection and segmentation challenge 2017 entry. In: ICCV COCO Challenge Workshop (2017)
13. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv:1511.00561 (2015)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Wang, J., Zhu, X., Gong, S., et al.: Attribute recognition by joint recurrent learning of context and correlation. In: IEEE International Conference on Computer Vision, pp. 531–540 (2017)
16. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep imbalanced attribute classification using visual attention aggregation. In: European Conference on Computer Vision (2018)
17. Liu, X., Zhao, H., Tian, M.: HydraPlus-Net: attentive deep features for pedestrian analysis. In: IEEE International Conference on Computer Vision, pp. 350–359 (2017)
18. Fang, W., Zhang, F., Sheng, V.S., Ding, Y.: A method for improving CNN-based image recognition using DCGAN. CMC: Comput. Mater. Continua **57**(1), 167–178 (2018)