Consumer Behavior Analysis in the Offline Retail Stores based on convolutional neural network

Mingxu Wang *
International School, Beijing University of Posts and Telecommunications, Beijing 100876, China

*Corresponding author e-mail: wangmingxu@bupt.edu.cn

Abstract. Pedestrian attributes recognizing (PAR) is an important task in computer vision area due to it plays an important role in video surveillance. On the other hand, pedestrian visual attributes are treated as middle-level semantic features which can provide calibration information for high-level human related visual tasks in order to improve the discriminative ability of the models, such as pedestrian detection, people tracking, person re-identification, action recognition and scene understanding. However, the most study of PAR is based on single person. In this work, I implements a multiple pedestrian attributes recognition model based on the offline retail scenes. This model combines object detection and multitask classification techniques and can be trained end-to-end directly for back propagation. This paper also demonstrates the performance of the final model through a series of experiment.

1. Introduction
Nowadays, most retail enterprises want to transform and upgrade the size of market by using the power of new technologies. The new retail has become a new enthusiasm for the future. The essence of the new retail is to reconstruct the "people-goods-field". But we have ignored shopping guides, another important component of "people", in the offline retail scene. We know that the closest to customers is shopping-guides. But in offline retail stores, especially in large chain retail stores, the phenomenon of misplacement of guided purchase or sales management, poor guided purchase, poor consumption experience, and low operational efficiency is widespread. In order to provide consumers with a better experience, there are many new retail solutions in the market, such as passenger flow analysis and statistics, accurate user portraits and other applications.

At the beginning, I achieved a network which predict pedestrian’s location and classifies the pedestrian’s attributes, which are male or female, guide or customer and pedestrian’s behaviors(standing or sitting, playing mobile phones or not playing mobile phones).

In order to get a better performance solution, I tried a variety of experiments. And this paper finally realized an end-to-end system which combine the object detection and the muti-task classification. The model is used to extract each pedestrian's location information (bounding box) and predict directly the classification of attributes(male or female, guide or customer) and behaviors (standing or sitting, playing mobile phones or not playing mobile phones) for every detected pedestrian.
At present, the research on pedestrian attribute recognition is mostly for single person. There is few end-to-end model which identify the pedestrian attributes on multi-person. In this case, I combine the object detection model and the multi-task classification model to train the model end-to-end.

2. Related Work

2.1. Foreign and domestic research situation
Pedestrian attributes are human-understandable semantic descriptions that distinguish with some traditional low-level features such as edges, textures, colors, HOG, LBP, Haar, etc. [1] Therefore pedestrian attributes are applied to many advanced computer vision tasks, such as person re-identification, face verification, and human identification. Pedestrian Attribute Recognition (PAR) is designed to predict the various attributes of a target person when given a person's image, as shown in Figure 1.

Traditional pedestrian attribute recognition adopts the manually methods to extracting more robust features and more powerful classifiers in order to get better performance. However, in the practical applications, the performance of these model is far below our expectations. With the development of deep learning, high-level semantic feature can be well extracted because of the accumulation large non-linear transformation. In turn, the direction of pedestrian attribute recognition has also made great progress.

However, the research on pedestrian attribute recognition and the data sets are mostly for single person. There is few end-to-end model which identify the pedestrian attributes on multi-person.

2.2. Deep Neural Network
Faster R-CNN, [2] published in NIPS 2015, is the third iteration and the first fully differentiated model of the R-CNN series models. Faster R-CNN consists of the following four parts:

1) Conv layers, extract the features of the image, input the whole image, and output the extracted features called feature maps.

2) Region Proposal Network (RPN), are used to proposal ROI. This network replace the previous search selective. The input is the feature maps extracted by conv layers [3] and the output is multiple bounding boxes and the probability of ‘foreground’ and ‘background’.

3) ROI pooling, convert different sizes of proposals into fixed-size feature map so that all proposal can be sent to the next classification network uniformly.

4) Classification and Boundary box regression. The output of this layer are the classes of the target object and bounding boxes of all foreground objects in the image.
2.3. Multi-task Classification
In the field of machine learning, for a particular task, we often design evaluation criteria, extract task-related high dimensional features and build a single or ensemble model. It uses high-dimension features and evaluation criteria to optimize model parameters to optimal performance, i.e., minimal loss. This pipeline will be satisfactory. However, the results of a single task ignore the possibility that other tasks can further improve the evaluation criteria.[4][7]

Therefore, the multi-task classification enables each sub-task to be mutually constrained and promoted together by joint training in order to achieve an overall minimum loss. Specifically, pedestrian attributes are interrelated, such as gender and identity (customer or guide). In supervised learning, the labeling often consumes a lot of manpower and financial resources.

Therefore, the most popular method is to joint learning multi-tasks to extract shared features so that we can use as little data as possible to get better performance of the models.[8] There are two ways of multi-task learning, i.e. the hard and soft parameter sharing, as shown in Figure 2.

3. Methods
3.1. Network Infrastructure
For our task, we need to use a multitasking classification model to classify multiple attributes of every pedestrian after the object detection model. Therefore, the common idea is to extract the position information of each pedestrian through the Faster R-CNN network, then crop each pedestrian and resize it into a unified size, and finally send into the multi-classification model in order to give each pedestrian's attributes labels. But, this pipeline separates the object detection model from the multi-classification model. Their training process is independent of each other and there is no any interaction of contextual information among them. Considering that the object detection task actually contains two tasks, that are bounding box regression and classification. Therefore, this paper uses VGG-net as the shared feature extraction network for object detection task and multi-task classification task. And then, modifies the classification task of the second stage in Faster R-CNN to multi-task classification loss, which includes object class label and multi-attribute labels.

The last is my multi-attribute classification module which contains six parallel small fully connection layers. And every every fc layers is responsible a pedestrian’s attributes, that is gender, customer, staff, stand, sit and play_with_phone. The architecture of the final network is shown in Figure 3.
Figure 3: the network architecture of multi-pedestrians attributes classification based on Faster R-CNN

The last is my multi-attribute classification module which contains six parallel small fully connection layers which achieve a two-classification task, as shown in Figure 13. And every every fc layers is responsible for predicting a pedestrian’s attributes, that is gender, customer, staff, stand, sit and play_with_phone.

4. Results
The final accuracy is compute by formula 1.

\[ P = P_0 \times 30\% + D \times 70\% \]  \hspace{1cm} (1)

Where, \( P_0 \) is positioning accuracy and \( D \) is pedestrian attributes accuracy.

The multiple pedestrian attributes recognition is achieved by combining the Faster R-CNN and the multi-classification. A demo of the result on test set is shown as Figure 4. And the test accuracy according to the evaluation criteria is 63.9461.
Extra experiment on pose estimation in order to help my model to predict the attributes that are stand, sit and play with phone. Pedestrian Attributes Recognition module is the main source of loss. Multitasking classification is achieved by sharing the weights of feature extraction network (Head) and the fully connection layers.[6] However, in order to prevent the explosion of parameters, I still design fc7 as a parameter sharing layer. This operation reduces the model parameters, but somewhat reduces the performance of the model, because the last fully connected layer containing 4096 neural sources is not sufficient to express the high-dimensional features for all classification tasks. In addition, the identity attribute (customer or staff) classifier mainly learns some color information of the wearer. Therefore, this classification may reduce the robustness of the model.

5. Conclusion
This paper uses computer vision, pattern recognition and other technologies to locate, identify, detect and analyze pedestrians' identities, attributes and behaviors, etc. Our system is mainly divided into three stages. First, pedestrian detection or target detection is used to extract each pedestrian's location information (bounding box). Secondly, the key points and posture features of each pedestrian are extracted by gesture recognition. The last stage is the classification of behaviors (standing or sitting, playing mobile phones or not playing mobile phones) by extracted posture features. In the process of implementation, because our training set is much smaller than the scale of CNN network used for image recognition, we need to apply migration learning technology. Retain the convolution layer and pooling layer of CNN network and training only the full connection layer, so as to avoid over-fitting.

References
[1] Xiao Wang, Shaofei Zheng, Rui Yang, Bin Luo and Jin Tang, (2019) “Pedestrian Attribute Recognition: A Survey,” arXiv preprint arXiv:1901.07474.
[2] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. International Conference on Neural Information Processing Systems.
[3] Zeiler, M. D. , & Fergus, R. . (2013). Visualizing and understanding convolutional networks.
[4] Rasmus, Antti, et al. (2015) "Semi-supervised learning with ladder networks." Advances in neural information processing systems.
[5] Chen X , Gupta A . (2017) An Implementation of Faster RCNN with Study for Region Sampling[J].
[6] Lin T Y , Goyal P , Girshick R , et al. (2017) Focal Loss for Dense Object Detection[J].
[7] IEEE Transactions on Pattern Analysis & Machine Intelligence, PP(99):2999-3007.

[8] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. JMLR, 11, 3371–3408.

[9] T. Evgeniou and M. Pontil. (2004) Regularized multi-task learning. In Proceeding of the tenth ACM SIGKDD international conference on Knowledge Discovery and DataMining.

[10] Ioffe S , Szegedy C . (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift[C]// International Conference on International Conference on Machine Learning. JMLR.3