HORA 2019
www.horacongress.com

International Congress on Human-Computer Interaction, Optimization and Robotic Applications
July 5-7, 2019, Urgup, Nevşehir, Turkey

SETSCI Conference Proceedings
www.set-science.com
8: 7 - 10, 2019

Real-time Pedestrian Attribute Detection from Surveillance Cameras

Betul Ay[1*] and Galip Aydin [2]

[1,2] *Computer Engineering, Fırat University, Turkey*
(betulay,galipaydin@firat.edu.tr)

Abstract – In recent years, computer vision has taken great strides in understanding and recognition the visual scenes together with deep learning technologies. Object recognition is one of the key areas of the computer vision applications. It is mainly concerned with recognition and localization of specific objects in an image. There are open-source and pre-trained models for the detection of general objects (such as cars, persons, cats, dogs). However, it is necessary to develop problem-based algorithms and training with deep neural networks by creating annotated training data for the recognition of special objects (such as headscarf). Furthermore, objects come in different shapes, sizes, angles, colors and additional noise from changes in the real-world environment, perspective, lighting and shadows. Taking into account these problems and needs in the real world, this paper focuses on the development of problem-based deep neural networks algorithms and the creation of labeled and reliable training datasets for the objects to be recognized. The contribution of the paper is to make use of transfer learning with the optimized R-FCN and Faster R-CNN pre-trained models in order to recognize pedestrian attributes including hat, headscarf, eyeglasses, bag objects and gender from security cameras. The proposed detection model has been trained on large-scale labeled dataset using TensorFlow open source platform. The performance of the neural network model has been evaluated using Average Precision (AP) values for each class and over 75% Mean Average Precision (mAP) for all classes is achieved.

Keywords – Object detection, deep learning, pedestrian attribute detection

I. INTRODUCTION

The recognition of pedestrian attributes such as gender, age, race [1], hat and bag detection from long range security cameras where the face is not detected or the facial features cannot be detected is a crucial task. This challenging task is performed using the fully body appearance and object detection methods. Object recognition has become an important field in autonomous vehicles that detect pedestrians, vehicles, traffic signs and lights, and in detecting objects such as people, bags, headscarves, glasses from long distance cameras. This paper addresses real-time pedestrian attribute detection on pedestrian images. Since the existing datasets (such as PETA) are insufficient and not suitable to recognize related attributes on whole images contains pedestrian, we collect our novel dataset contains an amalgamation of different datasets.

With the increase of surveillance cameras in recent years, intelligent video analysis technologies including object recognition and tracking have become compulsory. Many studies were conducted to recognize pedestrian attributes from far-view images [2-5]. Deng et al. [6] presents SVM-based method to inference the pedestrian attributes such as demographics (gender and age range), appearance (hair style), clothing style (casual or formal), and accessories (hat and backpack). For this task, they introduce a new dataset named as PETA consists of 19000 images with 61 annotated attributes. They have achieved average 71.1% classification accuracy with selected 35 attributes. Gupta and Ramesh [7] uses Convolutional Neural Networks (CNN) to recognize and detect pedestrian attributes on PETA dataset. Their neural network model has improved the attribute recognition performance with 80% test accuracy. They have proven that deep learning based approaches outperform traditional machine learning methods such as SVM.

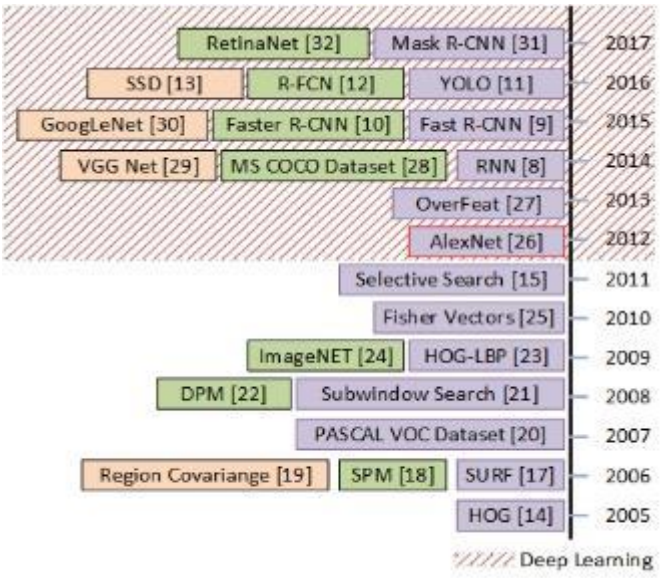

Fig. 1. Object detection timeline

Today, the success of object detection is based on the success of previous researchers. Recent progress has been made in the field of object recognition with these networks: R-CNN [8], Fast R-CNN [9], Faster R-CNN [10], YOLO [11], R-FCN [12] and SSD [13]. The timeline showing the development of object recognition between 2005 and 2018 is presented in Figure 1.

DOI: 10.36287/setsci.4.5.002

This progress began with a computer vision model called Histogram Oridient Gradients (HOG), which was first developed in 2005 by Navneet Dalal and Bill Triggs [14]. HOG features performed fast and well. However, with the increasing interest in deep learning, HOG characteristics performed worse and slower classification compared to the success and speed of CNN (Convolutional Neural Networks) architectures. For this reason, it has become important and prioritized to use the advantage of CNN architectures in object detection and image classification, to improve the classification capabilities and to increase the speed with selective search techniques used in particular for R-CNN [8]. Selective search [15] uses image properties to create all possible positions by looking at the pixel density, color, and image texture of an object. These identified objects are then exported to the CNN model for classification. In the following years, a model called Fast R-CNN [9], which corrects the main problems with R-CNN, has made better improvements in object recognition and classification. The Faster R-CNN model [10], which uses the small regional proposal CNNs instead of using selective search, provides a better way to solve this problem by providing better performance than the Fast R-CNN model. Faster R-CNN is one of the most accurate and fastest object detection algorithms in real-time use cases.

With the introduction of the R-FCN model in 2016 [12], more successful results were achieved in object recognition. The R-FCN model, which consists entirely of convolution layers, has achieved a faster and more accurate classification than the Faster R-CNN model, since it does not use any fully connected layers after the convolution layers. The R-FCN model, defined as an end-to-end network, directs the input received through the convolution layers directly to the output. The faster operation of end-to-end networks is predictable as it will have less weight than CNN models with a fully connected layer. For this reason, we use R-FCN model to pedestrian attribute detection in this paper.

The paper focuses on the use of problem-based deep neural network algorithms and the creation of labelled and reliable training sets for the objects to be recognized. It is mainly aimed to obtain the following 3 outputs:

- To provide a reliable data set by marking each object in the image with a label to identify the relevant object from the background.
- Real-time classification of various objects in the image.
- Localization to understand where objects are in the image.

II. Materials and Method

Human intelligence does not continuously learn new things from scratch, but instead transfers what they have learned in the past to new tasks. Transferring learning in artificial intelligence also works in this manner. It is a technique that allows us to reuse a previously trained (pre-trained) model for a new task by retraining the last layer of the model. In this paper, we build custom object detection model that classify and localize six pedestrian attributes (hat, headscarf, eyeglasses, bag, man, woman) by using a pre-trained model based on the Common Objects in Context (COCO) with the Tensorflow Object Detection API. We address the challenge of pedestrian attribute recognition problem with following deep learning architectures based on CNNs.

A. Region Proposal Networks

Region proposal is a list of bounding boxes that indicates likely positions of objects for given an input image. Region proposals are generated by Region Proposal Networks (RPN). R-CNN, Fast R-CNN and Faster R-CNN are the most widely known RPN approaches. R-CNN uses selective search to find region proposals per image and AlexNet to extract CNN features. For object classification and localization, Fast R-CNN uses a region of interest (RoI) layer unlike R-CNN. Each ROI is pooled into a fixed-size feature map and then the pooling layer mapped to a feature vector by fully connected layers. Finally, the network produces two output per RoI: softmax probabilities and bound box regression. While R-CNN and Fast R-CNN use selective search to find out region proposal, Faster R-CNN uses a separate network to predict the region proposals. R-FCN is also a RPN. Since fully connected layers after RoI pooling increase the complexity with the increase of the number of parameters, they are removed in R-FCN. R-FCN is faster that these approaches with competitive mAP [12].

B. Region-based Fully Convolutional Networks (R-FCN)

R-FCN is simple and efficient framework for object detection task and provides much faster training and inference performance, when compared with Faster R-CNN [12]. R-FCN is an end-to-end network: from input by convolutions to output. In other words, this architecture is composed of fully convolutional networks. Our pedestrian attribute detection system based on R-FCN is illustrated in Fig. 2.

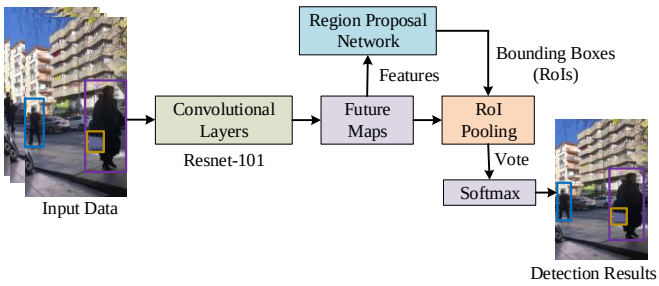

Fig. 2. Pedestrian attribute detection based on R-FCN

The network is started with fully convolutional layers and ended with RoI pooling layer. RPN generates candidate RoIs. The goal of the R-FCN architecture is to classify the RoIs into object categories and background. Before RoI pooling, the architecture uses Resnet-101 [33] to extract feature maps by removing the average pooling layer and the fully connected layer after convolutional layers. RoI pooling layer (named as position-sensitive RoI in [12]) takes the outputs of the last convolutional layer and produces scores for each RoI. Loss function on each RoI is computed as follows:

$$L(s, t_{x,y,w,h)} = L_{cls}(s_{c^*}) + \lambda[c^* > 0]L_{reg}(t, t^*) \quad (1)$$

Here, c is the number of object classes to be detected and c^* is the ground-truth label of RoIs. L_{cls} represents the classification loss and L_{reg} indicates the bounding box regression loss. t^* is the ground-truth box.

C. Dataset

Our large scale dataset, an amalgamation of different datasets with varying image sizes and resolutions, contains 1,120,020 labelled data. We split dataset into training and test sets in an 80–20 ratio. The dataset not only contains pedestrian images, but also includes images with tagged related attributes (hat, man, woman, eyeglasses, bag) taken from Google Open Images Dataset [16]. For headscarf object, we scrapped the related images from Google images and modanisa.com web site. We use LabelImg open source program to label object bounding boxes of our image data. This annotation tool saves an XML label for each image. After, we convert images and their xml files into a csv format for model training.

III. RESULTS

All the experiments were performed on the server which has 24-core Intel Xeon E5-2628L CPU, 8 NVidia GTX 1080-Ti GPUs, 256 GB RAM and runs Ubuntu Server 16.04 OS. TensorFlow platform and Python programming language has been used to train and test the neural network model. For training phase, we use Adam [34] optimizer and apply L2 regularization to the weights.

Evaluation metrics of object recognition are used to understand that how many objects were detected correctly as true positive and how many false positives were produced. We use Intersection over Union (IoU) to evaluate the object localization task and Mean Precision Average (mAP) to evaluate the detection task as evaluation metrics of the neural network model. This evaluation is performed based on ground-truth objects, which are annotated bound boxes drawn by human. IoU measures the overlap between two boundaries (ground truth and predicted). The measure is calculated by overlapping areas of intersection between these two boundaries. The output of this calculation is an accuracy score. We use IoU to measure how much our predicted boundary (the model prediction) overlaps with the ground truth. The mAP value ranges from 0 to 100 is calculated with the product of precision and recall of the detected bound boxes. If mAP value is above 0.5, the detection result is considered as true positive. For given ground-truth images, a high mAP value shows that the neural network model has high performance. Average precision (AP) for each classes and mean AP over these classes

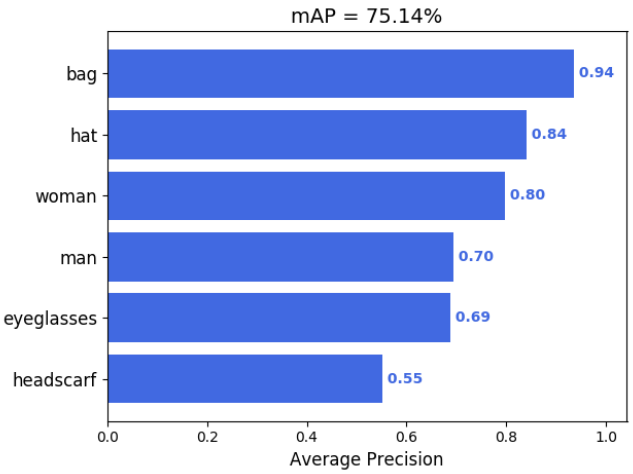

Fig. 3. Performance results for each object classes

(mAP) are reported in Fig. 3 to present the detection performance of the proposed model. Blue bars indicate AP values over each class for 1750 randomly selected test images. We manually annotated the ground-truth bound boxes of 1750 pedestrian images as shown in Fig. 3 (a). While the "bag" object has the highest AP value, "headscarf" has performed the worst with 55 % AP. Note that we do not have results from the baseline methods for "headscarf" because it is newly defined in this work. Ground-truth (a) and true-false positive predictions for each of the classes (b) are illustrated in Fig. 4. We observed that the most of the false positives for "man" objects are "woman" and vice versa. Although gender classification from facial features can be performed with very high accuracy rates, gender detection from a far-view camera where the face is not detected or the facial features cannot be detected is a difficult task. Our detection model has obtained average 75 % gender detection accuracy. We plan to further develop the model using larger and true annotated pedestrian images. We also observed that the model confuses human head with "headscarf" object. We make use of Python scripts (cloned from https://github.com/Cartucho/mAP) to show graphical detection results of the trained model. Fig. 5 shows that the example detection results of the model on the frames of a recorded video with a score threshold of 0.6.

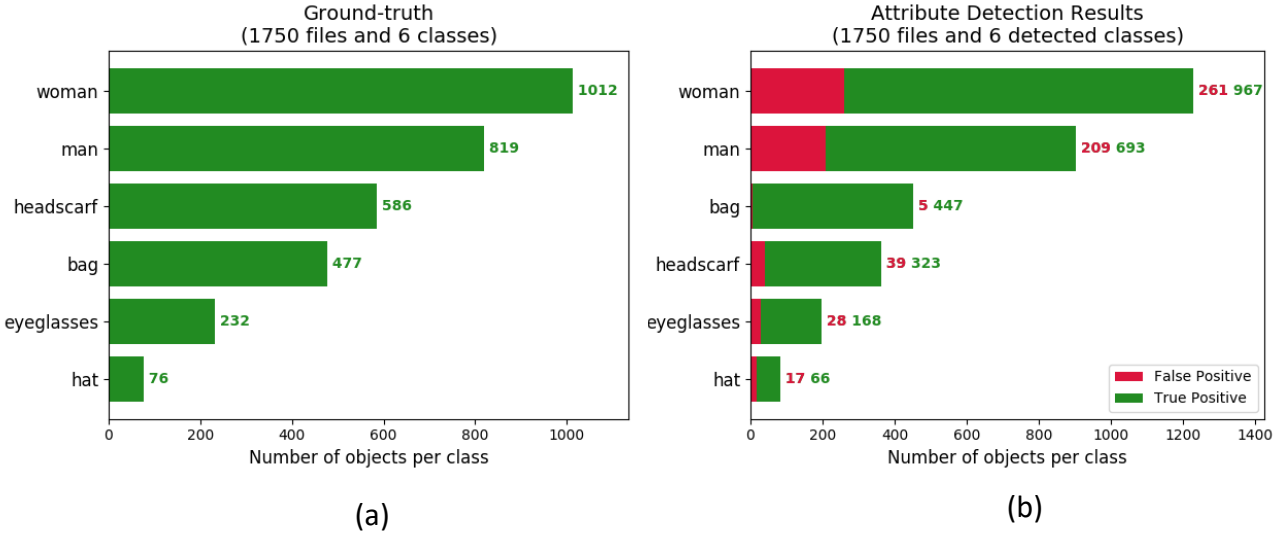

Fig.4. Ground-truth objects (a), attribute prediction results of the model with the number of false positive and true positive (b)

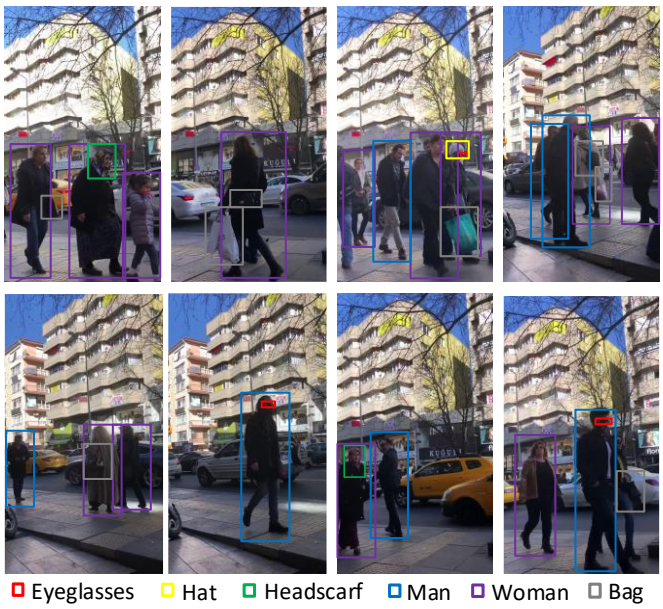

Fig. 5. Example results of the trained model on a video frame

IV. Conclusion

We trained and fine-tuned a pre-trained R-FCN model on our large dataset (~1M images) to detect six pedestrian attributes from far-view camera images. The model has efficiently performed in the real-time (the running time per image is 0.3sec on a computer with Intel Core i7 2.8 GHz CPU, 16 GB RAM. The computer also has a NVIDIA GTX 1070 GPU) and achieved high accuracy on test data. The main challenge in our study is the lack of labelled pedestrian images. We have just 20K pedestrian images taken from 20 different environments. So, we have combined our pedestrian images with Google Open Images Dataset. But Google Open Images Dataset contains images with too many false labels. As a result, we concluded that the model could get higher classification performance on accuracy with a larger dataset which contains more pedestrian labels.

Acknowledgment

This study was carried under the project “Deep Learning and Big Data Analysis Platform (DEĞİRMEN)” supported by Presidency of the Republic of Turkey, Presidency for Defence Industries (SSB).

References

[1] Talo, Muhammed, et al. "Bigailab-4race-50K: Race Classification with a New Benchmark Dataset." 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). IEEE, 2018.

[2] S. M. Li, Dangwei, et al. "A richly annotated dataset for pedestrian attribute recognition." arXiv preprint arXiv:1603.07054 (2016).

[3] Tian, Yonglong, et al. "Pedestrian detection aided by deep learning semantic tasks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[4] Zhu, Jianqing, et al. "Pedestrian attribute classification in surveillance: Database and evaluation." Proceedings of the IEEE international conference on computer vision workshops. 2013.

[5] Li, Dangwei, Xiaotang Chen, and Kaiqi Huang. "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios." 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, 2015.

[6] Deng, Yubin, et al. "Pedestrian attribute recognition at far distance." Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.

[7] Gupta, Agrim and Jayanth S. Ramesh. “Pedestrian Attribute Detection using CNN.” (2016).

[8] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[9] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[10] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[11] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[12] Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." Advances in neural information processing systems. 2016.

[13] Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.

[14] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." international Conference on computer vision & Pattern Recognition (CVPR'05). Vol. 1. IEEE Computer Society, 2005.

[15] Uijlings, Jasper RR, et al. "Selective search for object recognition." International journal of computer vision 104.2 (2013): 154-171.

[16] Kuznetsova, Alina, et al. "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale." arXiv preprint arXiv:1811.00982 (2018).

[17] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." European conference on computer vision. Springer, Berlin, Heidelberg, 2006.

[18] Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Vol. 2. IEEE, 2006.

[19] Tuzel, Oncel, Fatih Porikli, and Peter Meer. "Region covariance: A fast descriptor for detection and classification." European conference on computer vision. Springer, Berlin, Heidelberg, 2006.

[20] Everingham, Mark, et al. "The PASCAL visual object classes challenge 2007 (VOC2007) results." (2007).

[21] Lampert, Christoph H., Matthew B. Blaschko, and Thomas Hofmann. "Beyond sliding windows: Object localization by efficient subwindow search." 2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008.

[22] Felzenszwalb, Pedro F., et al. "Object detection with discriminatively trained part-based models." IEEE transactions on pattern analysis and machine intelligence 32.9 (2009): 1627-1645.

[23] Wang, Xiaoyu, Tony X. Han, and Shuicheng Yan. "An HOG-LBP human detector with partial occlusion handling." 2009 IEEE 12th international conference on computer vision. IEEE, 2009.

[24] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[25] Csurka, Gabriela, and Florent Perronnin. "Fisher vectors: Beyond bag-of-visual-words image representations." International Conference on Computer Vision, Imaging and Computer Graphics. Springer, Berlin, Heidelberg, 2010.

[26] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[27] Sermanet, Pierre, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229 (2013).

[28] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[29] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[30] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[31] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[32] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[33] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[34] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).