An object Detection System Based on YOLOv2 in Fashion Apparel

Zhihua Feng
School of Computer Science & Technology,
Donghua University
Shanghai, China
e-mail: fengzhihua2012cc@163.com

Xin Luo
School of Computer Science & Technology,
Donghua University
Shanghai, China
e-mail: xluo@dhu.edu.cn

Tao Yang
Donghua University
Shanghai, China
e-mail: yangtao@dhu.edu.cn

Kenji Kita
Faculty of Engineering, Tokushima University
Tokushima, Japan
e-mail: kita@is.tokushima-u.ac.jp

***Abstract*—Object detection of fashion apparel items is an indispensable technical condition for fashion industry and e-commerce. To address this need, this paper constructs an object detection system (YOLOv2-opt) based on the YOLOv2. In this work, we innovatively combine deep convolutional neural network with fashion apparel items detection. Compared with the traditional machine learning algorithms and semantic segmentation, the method based on deep neural network can locate and classify objects fast and better. Since the characteristics of fashion apparel detection, we propose an optimization for the input and network layers of YOLOv2 in order to improve the detection performance. The types of fashion items we consider in this work include trousers, skirts, coats, T-shirts, bags. Through the experiments, we evidence that our system has achieved an average precision of 0.839 and average recall rate of 0.73, detection speed reached 56ms per picture. Our optimized system is more exact than the YOLOv2.**

Keywords-component; object detection; deep learning; fashion apparel

I. INTRODUCTION

Fashion apparel in social life reflects age, social status, lifestyle and gender. Clothing and apparel play an important role in life. Clothing is an important descriptor for recognizing human beings, such as “man wearing a jacket” or “a woman wearing high heels”. "Fashion Apparel Detection" also has many applications. For example, marking the details of clothing in unmarked images and applying them to a mass e-commerce database to achieve a simple classification search. Similarly, the different fashion clothing detected in one image can be regarded as a style match, so that the training can be recommended by the fashionistas after the dressing match. Real-time clothing detection can detect the presence of dangerous signs in the environment, such as gangster type dress or deliberately exposed. In this paper, we treat fashion apparel detection as “Algorithm capable of taking a photograph of a person wearing fashion apparel as input, and discovering the various fashion apparel present in the image”.

At present, most object detection processes [1] are mainly divided into three steps: the first step is to input an image and divide the target object into regions. In the second step, image features are extracted and image feature are used to represent the object. In the third step, according to the feature classification or image semantic segmentation, the area is defined and the object is framed to achieve the object detection. The traditional detection methods have the disadvantages of slow speed, low accuracy and complicated process, which can't reach the idea of quickly detecting the fashion clothing items in this paper.

Object detection based on deep learning is superior to traditional feature extraction and classifier detection methods in terms of speed and accuracy. The core of the deep learning method is the convolutional neural network, which is a forward feedback neural network, and has a unique advantage in object recognition and has a special weight distribution structure. CNN (Convolution neural network) mainly includes convolution layer, pooling layer and fully-connected layer. Object detection based on CNN can be simply divided into two parts: classification and position. Some region-based object detection algorithms, such as R-CNN, Fast-RCNN [2], etc., achieve classification by CNN while extracting region proposals with the method selective search. Extracting region proposals spends a relatively long time. YOLO, SSD [3] proposed to fit the position and classification information of the object in a CNN to form an end-to-end object detection model. YOLO is a fast and efficient object detection algorithm, which is proposed by Joseph Redmon [4]. YOLO can detect 45 images per second, while YOLOv2 can detect single images at speeds up to 11ms.

II. RELATED WORK

The first segmentation-based fashion spotting algorithm for general fashion items was proposed by [5] where they introduce the Fashionista Dataset and utilize a combination of local features and pose estimation to perform semantic segmentation of a fashion image. In [6], the same authors followed up this work by augmenting the existing approach

978-1-5386-8339-2/18/$31.00 ©2018 IEEE

with data driven model learning, where a model for semantic segmentation was learned only from nearest neighbor images from an external database.

Apart from the above two works, [7] also proposed a segmentation-based approach aimed at assigning a unique label from “Shirt”, “Jacket”, “Tie” and “Face and skin” classes to each pixel in the image. Their method is focused on people wearing suits.

With the continuous development of the clothing industry, people began to try to apply some methods of deep learning to fashion clothing detection. In the literature [8], Brain Lao et al. used the fine-tuned R-CNN model to detect objects in fashion apparel under the caffe. Under the pre-defined "shoes" and "belt" categories, 91.25% of the highest verification accuracy was obtained. Their model has a good effect in detecting the "shirts", but it is not effective in detecting the "belts", which are not obvious and easy to mix with background information.

Kota Hara et al. [9] proposed a deep convolutional network algorithm that combines background information of body postures to solve the dependency and deformability of fashion clothing. They consider the variability of fashion clothing, but the position of the object to be detected is closely related to the posture of the human body. Therefore, they proposed a model that relies on posture detection, which can provide more accuracy for fashion clothing. But at the same times it increases the detection time, the algorithm can process a picture for up to 10s, which can’t achieve the effect of real-time detection.

III. PROPOSED METHOD

The purpose of this system is to quickly detect and label fashion clothing in a given image. Based on the YOLOv2[10] object detection algorithm, this paper constructs an object detection system (YOLOv2-opt) and optimizes the entire detection process. Firstly, the input data is preprocessed to achieve the effect of data enhancement, which is used to improve the detection accuracy of the algorithm. Secondly, we change the maxpooling layer of the original convolutional neural network to S3Pool, which improves the generalization ability of image data, thereby improving the detection accuracy of fashion clothing detection. Figure 1 is a schematic diagram of the object detection system.

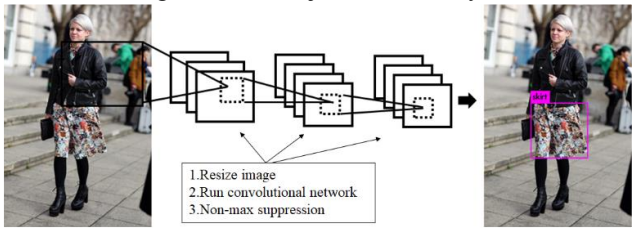

Figure 1. Processing images with object detection system.

A. Object Proposal

The YOLOv2 method treats object detection as a regression problem, and a convolutional neural network structure can directly predict the bounding box and classification from the input image. The YOLOv2 first divides the input picture into S×S grids, each of which is responsible for detecting objects falling in the grid, and each grid returns the bounding box information and the probability belonging to a certain classification.

Each bounding box[11] consists of 5 predictions: x, y, w, h, and confidence. The x, y coordinates represent the center of the box relative to the bounds of the grid cell. The w, h are predicted relative to the whole image. Finally, the confidence prediction is represented by the IOU between the prediction box and any ground truth boxes. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts. The confidence formula is:

$$\text{Confidence} = \Pr(\text{Object}) * \text{IOU}_{pred}^{truth} \quad (1)$$

IOU is the ratio of the prediction box to the real box:

$$IOU_{Pred}^{Truth} = \frac{area(box(Truth) \cap box(Pred))}{area(box(Truth) \cup box(Pred))} \quad (2)$$

Pr(Object) is the probability that the bounding box contains the object. box(Truth) reference ground truth box based on training tags, box(Pred) reference prediction box, area() indicates area.

It divides the image into an S × S grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an S × S × (B×5 + C) tensor.

B. Network Structure

YOLOv2 uses a new classification network as feature extraction. The author designed the network model Darknet-19 with reference to the advantages and disadvantages of the current classification network, such as VGG. Darknet-19 has 19 convolutional layers and 5 maxpooling layers. For a full description see Table Ⅰ. The network uses a global average pooling and places a 1×1 convolution kernel between the 3×3 convolution kernels to compress features. Adding Batch Normalization between layers improves the convergence speed of the model and solves the gradient problem in back propagation. The network removes the fully connected layer and uses the convolution layer to detect the confidence of the bounding box. This approach draws on the key steps of the RNP network, performing a sliding frame operation on the convolutional layer, and selecting nine different sizes of anchor boxes from a center. This calibrates the spatial information of the picture more than the full connection layer coordinate positioning. The anchor boxes are trained by the K-means clustering method class. The scoring criteria are determined by the IOU, so the distance function can be written as:

$$min \sum_{N} \sum_{M} (1 - IOU(Box[N], Truth[M]))$$

Fashion clothing detection mainly considers five categories of trouser, skirts, jackets, T-shirts, and handbags. This article changes the class probability of the output layer to 5. That is, S=7, C=5.

TABLE I. NETWORK

layer name	Filter	Size/Stride	Output
Convolutional	32	3×3	224×224
Maxpool		2×2/2	112×112
Convolutional	64	3×3	112×112
Maxpool		2×2/2	56×56
Convolutional	128	3×3	56×56
Convolutional	64	1×1	56×56
Convolutional	128	3×3	56×56
Maxpool		2×2/2	28×28
Convolutional	256	3×3	28×28
Convolutional	128	1×1	28×28
Convolutional	256	3×3	28×28
Maxpool		2×2/2	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Maxpool		2×2/2	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	1000	1×1	7×7
Avgpool		Global	1000
Softmax			

C. *Accuracy Optimization*

In this paper, according to the characteristics of fashion clothing pictures, the training pictures are preprocessed, so that the human body and clothing features in the pictures are more prominent. Changing the pooling layer method to improve the image's generalization and reduce the loss of image features. Considering the diversity of fashion clothing styles and the characteristics of easy deformability, this paper adjusts the proportion of classification loss in the loss function to better fit the classification.

1) Pre-processing：Perform some pre-processing on the image before object detection[12], which can effectively increase the detection accuracy. Since the experimental method of this experiment is aimed at the street pedestrians' costumes, the combination of pedestrians and street background tends to make the picture partially dim or partial highlights. Therefore, we use the preprocessing method based on histogram equalization to process the picture. Histogram equalization is a histogram of known grayscale probability density distribution images that are modified by a transformation function to have a uniformly distributed histogram. Histogram equalization can improve the overall brightness and gray level of the image, so that the contrast of the image increases, and the character of the figure in the image is more obvious. This paper adjusts the parameters to make the input image reach an optimal state. By preprocessing the data set by this method, the object detection accuracy can be improved by 2%-3%.

2) S3Pool[13]: YOLOv2 uses the maxpooling to downsample the feature map to reduce the spatial dimension of the feature map. While the typical downsampling step of a maxpooling layer intuitively reduces the spatial dimension of a feature map by always selecting the activations at fixed locations, this design choice is somewhat arbitrary and potentially suboptimal. If we allow the downsampling step to be performed in a non-uniform and non-deterministic way, where the sampled indices are not restricted to be at evenly distributed locations, we are able to produce many variations of downsampled feature maps. We use S3Pool to achieve these.

S3Pool regards the maxpooling as two steps. In the first step, the pooling window of the K×K size performs a maxpooling of the image matrix in a step size of 1. In the second step, the spatial down sampling step is performed, and the upper left corner element in the S×S (original step size) window is selected as the output feature map. S3Pool uses the same first step as the maxpooling, and the spatial matrix is divided by the gradient g before downsampling the spatial features. The gradient g controls the amount of distortion introduced in the downsampling step so that the regularization intensity of the feature extraction becomes controllable. The final feature is chosen as the expected value in the downsampling process in the g×g window. The comparison between the maxpooling and the S3Pool method is shown in Figure 2.

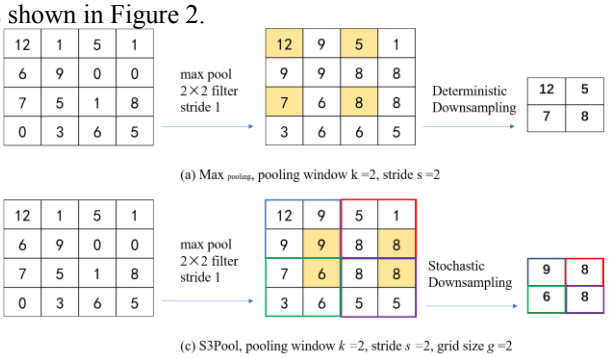

Figure 2. Comparison of different pooling methods (best seen in color).

3) Fit loss function: the loss function of the YOLOv2 is to achieve a good balance between the coordinates (x, y, w, h), confidence, and classification. That is, the mean squared error of the S × S (B × 5 + C) dimensional vector of the network output and the corresponding S × S (B × 5 + C) dimensional vector of the real image. This paper attempts to change the proportion of classification in the loss function[14]. Taking into account the large size of the fashion clothing box but not easy to classify, the classification ratio is raised to a reasonable proportion.

As shown in the following formula. Where I_i indicates whether an object falls into the i-grid, yes to 1, not to 0. I_{ij}

indicates whether the j_{th} detection frame in the grid i is responsible for the object, yes to 1, not to 0. The first two lines of the formula are responsible for predicting the coordinates, the middle two are the calibration frame confidence prediction, and the last line is responsible for the prediction category. This paper adds a hyperparameter λ_{class} to adjust the role of classification prediction in the entire loss function.

$$\lambda coord \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]$$
$$+ \lambda coord \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$
$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left(C_i - \hat{C}_i \right)^2$$
$$+ \lambda noobj \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} \left(C_i - \hat{C}_i \right)^2$$
$$+ \lambda class \sum_{i=0}^{S^2} I_{ij}^{obj} \sum_{c \in class} (p_i(c) - \hat{p}_i(c))^2$$

IV. Experimental Results and Discussion

The experimental environment of this article is ubuntu14.04 operating system, compile Darknet and configure cudnn acceleration. We used Nvidia Tesla K80 GPU to accelerate training and testing of data sets.

A. Data Set

This paper applies an open source set of streets beat data sets CCP [15] for training and testing, a total of two thousand images with a resolution of 550×830. Due to the high image quality and rich color content of this data set, it is more in line with the purpose of detecting the fashion clothing in the complex background image, and simulating the effect of detecting in a monitoring system. We perform calibration including classification and bounding box on the data set. In the training set, the calibration is divided into pants, skirts, jackets, handbags and T-shirts. The number of marked pictures is 1,800. The total number of samples marked is 4605. 200 pictures are used for testing.

According to the VOC directory structure, firstly, we convert the annotation of our data set into XML format. Files' paths of training set and testing set are written to train.txt and test.txt. It is convenient for training process.

B. Training Process

We conducted a contrast experiment with the optimized system constructed in this paper, YOLOv2-opt and YOLOv2, to prove that the optimization proposed in this paper is effective. In this paper, the prepared training data sets are placed in the above algorithm for training. The number of iterations in this paper is set to 42000, and the momentum and weight attenuation are 0.9 and 0.005. The initial learning rate was set to 0.0001, which was adjusted to 0.001, 0.0001, and 0.00001 at 100, 25000, and 35,000 times, respectively.

This paper uses the pre-training model darknet19.conv.23 for accelerated training. Using our training set, training is carried out about 2 days with GPU. Then the final object detection weight model is obtained.

C. Result analysis

Analysis of training results, the average detection speed of the YOLOv2-opt object detection system and other object detection algorithms are compared. The comparison shows that the YOLOv2-opt algorithm is generally faster than other algorithms. YOLOv2-opt takes 56ms to process an image, which is basically the same as the YOLOv2 speed before it was improved. This paper compares the speed of a fashion clothing picture detected by YOLOv2 and YOLOv2-opt with the baseline speed of other object detection algorithms in Figure 3.

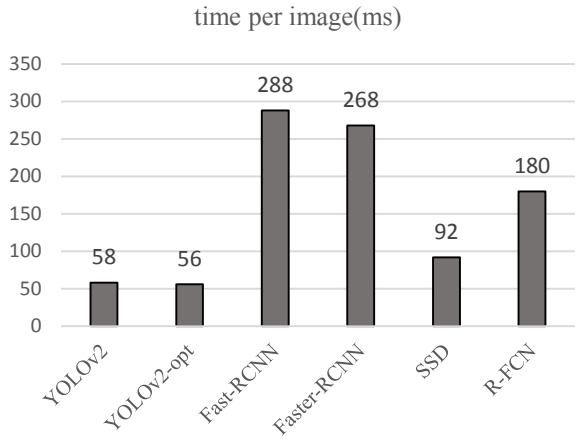

Figure 3. Speed of each detection algorithm.

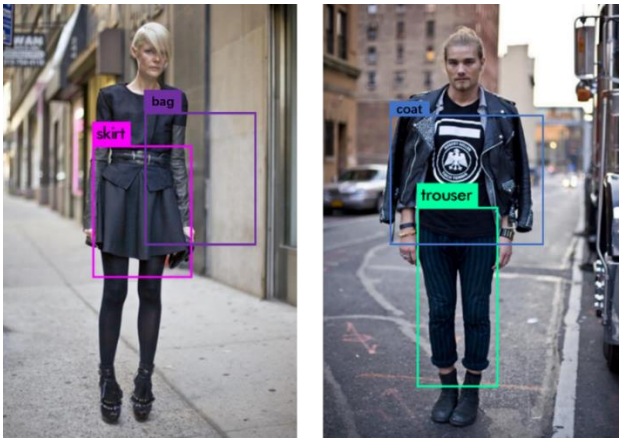

Figure 4. Object detection results of images

The single picture running the YOLOv2-opt model gets the effect shown in Figure 4. It is not difficult to find that in the fashion clothing detection, such as trouser and skirt with

relatively simple structure have a relatively accurate recognition rate. In the female figure on the left, the dark background is mistakenly recognized as a bag, indicating that the model needs to be enhanced in processing complex images. In this paper, the recall and the precision [16] are used to evaluate the quality of the training model, and the images of 200 testing sets are used for testing.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

TP, FP, and FN are True Positive, False Positive, and True Negative. Calculate the IOU value of the detection bounding box and the reference bounding box to determine the result of TP, FP and FN. IOU$\geqslant$0.5, TP. IOU<0.5, FP. IOU = 0, FN. The results are shown in Table II. Our YOLOv2-opt model for fashion apparel has a precision of 0.839 and a recall rate of 0.73, higher than the YOLOv2 model.

TABLE II. RESULTS

$YOLOv2-opt/_{YOLOv2}$	*Total*	*Total of Detection*	*TP*	*FP*	*Precision*	*Recall*
trousers	95	92/88	86/79	6/9	0.935/0.898	0.905/0.832
skirt	62	58/55	54/48	4/7	0.931/0.873	0.871/0.774
coat	126	96/82	72/56	24/26	0.75/0.683	0.571/0.444
T-shirt	46	38/42	28/26	10/16	0.737/0.619	0.609/0.565
bag	64	58/48	47/39	11/9	0.81/0.813	0.734/0.609
Total	393	342/315	287/248	55/67	0.839/0.787	0.73/0.631

V. CONCLUSION

In this work, our proposed optimization system (YOLOv2-opt) improves the accuracy of fashion apparel object detection. The system basically solves the problem of quickly discovering various fashion Apparel in the image. At the same time, although the data set is representative, but the overall number is too small. The next study can consider limiting the class position of fashion clothing in the picture in accordance with more auxiliary information, such as human body information, head information, environmental information, etc., to improve the positioning.

ACKNOWLEDGMENT

The authors would like to thank Guangdong Province collaborative innovation and platform Environmental Science build of special funds (2014B090908004); Dongguan City professional town innovation service platform construction project" Dongguan City Humen garment Collaborative Innovation Center", whose constructive comments and suggestions helps us to improve the quality of this paper.

REFERENCES

[1] Yamaguchi K. Parsing clothing in fashion photographs[C]// Computer Vision and Pattern Recognition. IEEE, 2012:3570-3577.

[2] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]// Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2003:I-511-I-518 vol.1.

[3] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision. Springer International Publishing, 2016:21-37.

[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[5] Yamaguchi K, Kiapour M H, Ortiz L E, et al. Parsing clothing in fashion photographs[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012:3570-3577..

[6] Yamaguchi K, Kiapour M H, Berg T L. Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items[C]// IEEE International Conference on Computer Vision. IEEE Computer Society, 2013:3519-3526.Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014:580-587.

[7] Hasan B, Hogg D. Segmentation using Deformable Spatial Priors with Application to Clothing[C]// British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings. DBLP, 2013:1-11.

[8] Lao B, Jagadeesh K. Convolutional neural networks for fashion classification and object detection[J]. 2015-09-07 J. http://es231 n. stanford,edu/reports/BLAO_KJAG_CS231N_FinalPaperFashionClassifieation. pdf, 2016.

[9] Hara K, Jagadeesh V, Piramuthu R. Fashion apparel detection: the role of deep convolutional neural network and pose-dependent priors[C]//Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016: 1-9.

[10] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[J]. 2016:6517-6525.

[11] Redmon J, Angelova A. Real-Time Grasp Detection Using Convolutional Neural Networks[J]. 2014, 2015:1316-1322.

[12] Gao J, Xue-Wei L I, Zhang J. Study of Image Pretreatment Algorithm of Quality Detection System for Color Printing[J]. Packaging Engineering, 2007, 9(3):259-268.

[13] Zhai S, Wu H, Kumar A, et al. S3Pool: Pooling with Stochastic Spatial Sampling[C]// Computer Vision and Pattern Recognition. IEEE, 2017:4003-4011.

[14] Ko Y H, Kim K J, Jun C H. A New Loss Function-Based Method for Multiresponse Optimization[J]. Journal of Quality Technology, 2005, 37(1):págs. 50-59.

[15] Yang W, Luo P, Lin L. Clothing Co-parsing by Joint Image Segmentation and Labeling[C]// Computer Vision and Pattern Recognition. IEEE, 2014:3182 - 3189.

[16] Wang Wei. Full recall and precision rate[J]. Information Science, 1981(3): 40-44.