Smart Phone Aided Intelligent Invoice Reimbursement System

Yang Meng, Yan Liang, Yingyi Sun, Jinqiu Pan, and Guan Gui(✉)

College of Telecommunications and Information Engineering,
Nanjing University of Posts and Telecommunications, Nanjing 210003, China
guiguan@njupt.edu.cn

Abstract. Invoice reimbursement is one of indispensable apsects of business in many countries especially in China. Conventional manpower based reimbursement schemes often lead to high cost and inefficiency and robot based reimbursement systems require large space and huge equipment costs. In order to solve these problems, we propose an smart phone aided reimbursement system to realize the intelligent localization and identification in invoice images. First, invoice image is taken by camera of smart phone. Second, the Hough transform is used to detect the linear principle to correct the tilt of the invoice image with different background and different tilt angles. Third, we adopt You Only Look Once-Version 3 (YOLOv3) based target detection network to train the tagged data set, to obtain the training weights, and then realize the intelligent positioning and extraction. Finally, the invoice information is identified using optical character recognition (OCR). Experiment results are given to verify that the localization accuracy can reach 92.5% when the intersection over union (IoU) is set as 0.5 and the identification accuracy can reach up to 97.5% for invoice information.

Keywords: Deep learning · Intelligent positioning · Hough transform · Optical character recognition (OCR) · Invoice information identification

1 Introduction

With the rapid development of the economy, invoices are very popular for business transactions and consumption reception [1]. Invoice reimbursement problem poses a big challenge for many companies and governments. Invoice reimbursement is one of indispensable apsects of business in many countries especially in China. At this stage, The entire reimbursement process relies mainly on manpower in China. The entire process of reimbursing invoices is extremely cumbersome and complicated, and while consuming a large amount of manpower, it results in low reimbursement efficiency and high error rates. The traditional invoice image localization method is implemented for the position coordinates of the target area. The specific method is to preprocess the irregular image, and find the coordinates corresponding to the key information area on the invoice according to the size of the whole picture. For the pixels of different pixels and different sizes to be re-edited, it is not universal, and there is no way to achieve intelligent positioning. Another relatively intelligent method is the template matching

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 302, pp. 320–329, 2019.
https://doi.org/10.1007/978-3-030-36405-2_32

method [2], which is to find a specific target in an image. The principle is very simple. It traverses every possible position in the image and compares it with the template. When the similarity is high enough, it is considered that the correct target is found. Although the template matching method can achieve the purpose of precise intelligent positioning, it has high requirements on the image to be positioned, and must be an image with no background interference and the same resolution. This method is also not universal. None of the above methods can achieve intelligent positioning in the true sense.

The system realizes intelligent positioning and content recognition of invoice images taken by mobile terminals with different backgrounds, different sizes, different resolutions and different angles, and the recognition accuracy is very high. The realization of the function of this system mainly focuses on the following aspects. First, because the paper invoice is easily damaged and lost, resulting in the problem of unable reimbursable, the system solves this problem by intelligently locating the invoice image taken by the mobile terminal. Second, the OCR is used to identify the entire invoice image, and the result of the recognition is very confusing, and the key information corresponding to the key segment cannot be found.

The system divides the invoice image into several areas. After intelligently locating different areas, according to the returned bounding box coordinates, the OpenImage Computer Vision Library (OpenCV) [3] CropImage image cropping tool is used to intelligently crop the positioned area and then use OCR to identify [4]. The accuracy is greatly improved. Third, the premise of traditional positioning is that the image being positioned has no background interference and high resolution requirements. The system uses the Hough transform detection linear principle, only need to tilt the invoice image correction, there is no requirement for the image background part, resolution and tilt angle, and it has universality. Fourth, the traditional positioning function has limitations, and it is impossible to intelligently locate images with different backgrounds, different resolutions, and different tilt angles. The system is based on deep learning, and uses the YOLOv3 target detection network to perform key areas in the invoice image. Feature extraction, which realizes the intelligent positioning of the invoice image captured by the mobile terminal. This system has a profound impact on the field of deep learning and computer vision image processing.

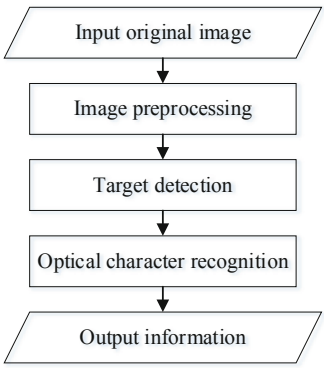


Fig. 1. Flow chart of the entire process

The flow chart of the system is provided in Fig. 1. The remainder of this paper is arranged as follows. In Sect. 2, the process of the proposed method is explained. Section 3 describes each specific step in detail and experimental results. We conclude the paper in Sect. 4.

2 System Design

We propose to apply the Hough transform detection linear principle to the preprocessing of invoice images. Then we use the YOLOv3 target detection network to intelligently locate and extract invoice images taken in natural scenes. Finally, we use the weak supervised learning framework to perform accurate character recognition on the extracted invoice image.

2.1 Image Preprocessing

The invoice image is mainly composed of straight lines, and the principle of Hough transform detecting lines can be applied to this aspect. The invoice image is mainly composed of straight lines, and the principle of Hough transform detecting lines can be applied to this aspect. The angle of the tilt in the principle of the straight line is detected, and the entire invoice image is corrected by a certain angle rotation.

The Hough transform uses a transformation between two coordinate spaces. A curve or line having the same shape in one image is mapped to a point on another coordinate space to form a peak, and then a problem of detecting an arbitrary shape is converted into a statistical peak. Using the principle of Hough transform to detect straight lines, we apply it to the intelligent correction of invoice images taken at different angles on the mobile end, and the effect is very good. The Hough transform detection linear principle polar coordinate diagram is shown in Fig. 2.

If the edge of the invoice is broken, the principle of Hough transform detects the line can also intelligently correct the damaged invoice image. This method is innovatively applied to the field of invoice image intelligent correction, effectively solving the problem that the traditional method cannot correct the incomplete image of the edge damage. The main reason is that this method can completely ignore the background and only perform the function of intelligently correcting the image for the straight line on the invoice image.

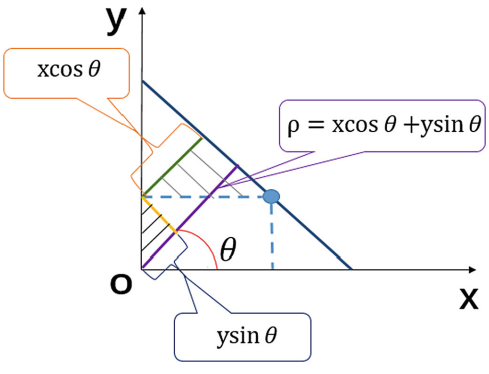
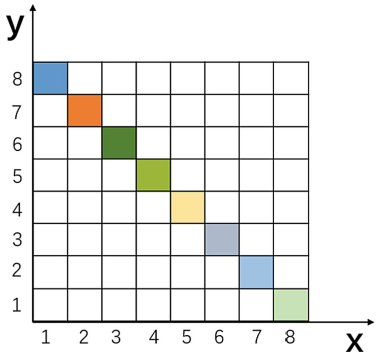


Fig. 2. Hough transform detection linear principle polar coordinate diagram.

For the better calculations, the Hesse normal form is proposed with the following formula:

$$\rho = x\cos\theta + y\sin\theta \tag{1}$$

where ρ is the distance from the origin to the nearest point on the line, and θ is the angle between the x-axis and the line connecting the origin and the nearest point. Each line of the image can be associated with a pair of parameters (ρ, θ), which is called a Hough space. The schematic diagram is shown in Fig. 3.



ρ \ θ	0°	45°	90°	135°	180°
$1\cos\theta + 8\sin\theta$	1	$9\sqrt{2}/2$	8	$7\sqrt{2}/2$	-1
$3\cos\theta + 6\sin\theta$	3	$9\sqrt{2}/2$	6	$3\sqrt{2}/2$	-3
$5\cos\theta + 4\sin\theta$	5	$9\sqrt{2}/2$	4	$-\sqrt{2}/2$	-5
$7\cos\theta + 2\sin\theta$	7	$9\sqrt{2}/2$	2	$-5\sqrt{2}/2$	-7
$8\cos\theta + 1\sin\theta$	8	$9\sqrt{2}/2$	1	$-7\sqrt{2}/2$	-8

Fig. 3. Hough transform voting algorithm.

Suppose there is a straight line in an 8×8 plane pixel, and the coordinates (x, y) corresponding to the eight Descartes coordinate systems are converted into polar coordinates (ρ, θ). When θ is taken at different angles, the value of ρ is obtained. Calculate all ρ values, the maximum number of votes is ρ. As shown in the figure above, the equation of the line is $9\sqrt{2}/2 = x\cos 45^o + y\sin 45^o$.

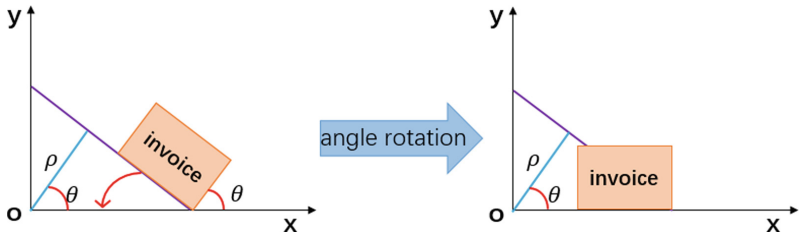


Fig. 4. Tilt correction of invoice image.

Using the detected straight line in the Hough transform, the tilt angle of the invoice can be calculated, and then the invoice can be corrected by rotating a certain tilt angle. This part is embedded in the test script of YOLOv3 to complete the correction of the invoice image taken at different angles of the mobile terminal. The tilt correction results are shown in Fig. 4.

2.2 Target Detection

YOLOv3 belongs to the target detection method of the regression sequence. Unlike the detection method of the sliding window and the subsequent area division, it regards the detection task as a regression problem and uses the neural network to directly predict the coordinates of the bounding box from the entire image. The box contains the confidence of the object and the class probability of the object, which can achieve end-to-end detection performance optimization. YOLOv3 detects objects very fast, about 45-155FPS, and YOLOv3 can avoid background errors and generate false positives. Because of this, we use YOLOv3 to extract features of invoice images with different backgrounds, different angles, and different pixels to achieve intelligent positioning and content extraction.

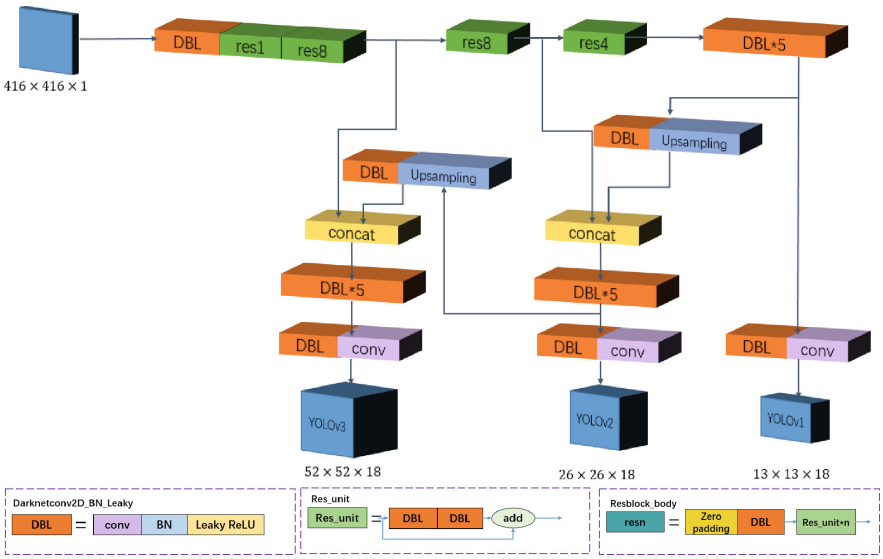


Fig. 5. Yolov3 network architecture.

As the latest algorithm in the YOLO series [5], YOLOv3 has both reservations and improvements to the previous algorithms [6]. The network architecture of YOLOv3 is shown in Fig. 5. Starting with YOLO, the YOLO algorithm does the detection by dividing the cells, but the number of divisions is different. Use "Leaky ReLU" as the activation function. A loss function completes the training, focusing on the input and output to achieve end-to-end training. Starting with YOLO 9000 [7], YOLO uses batch normalization as a method of regularization, accelerated convergence, and avoiding overfitting, connecting the BN layer and the Leaky ReLU layer to each layer of the convolutional layer. There is a choice between speed and accuracy. If you want to train fast, you can sacrifice accuracy. If you want high accuracy, you can sacrifice a little speed to achieve multi-scale training.

The improvement of each generation of YOLO depends largely on the improvement of the backbone network. From darknet-19 of YOLO9000 to darknet-53 of YOLOv3. YOLOv3 also offers alternate use of Darknet-53 and tiny-darknet. To improve performance, Backbone can use Darknet-53; for lightweight and high speed, you can use tiny-darknet.

2.3 Natural Scene Text Detection Technology

After completing the intelligent positioning and extraction of the invoice image taken by the mobile terminal, the OCR technology of deep learning is used to identify the text content of the invoice.

The text recognition of the invoice image is divided into two specific steps, the detection of the text and the recognition of the text, both of which are indispensable. Text detection of invoice images is very challenging when invoices exist in complex scenarios. Text detection in natural scenes has the following difficulties: text has multiple distributions; text layout is diverse; text has multiple directions; multiple languages are mixed.

At present, there are several popular text detection technologies based on natural scenes: the Connectionist Text Proposal Network (CTPN) network architecture commonly used in the OCR system proposed in 2016 [8]; the SegLink deep learning neural network proposed in 2017 that incorporates the small-scale candidate box of CTPN and learns from the Single Shot Multi Box Detector (SSD) algorithm [9]; the end-to-end text detection An Efficient and Accurate Scene Text Detector network architecture (EAST) [10] proposed in 2017; the weakly supervised Exploiting Word Annotations for Character based Text Detection network architecture proposed by Baidu in 2017 [11].

In this paper we use the weakly supervise exploiting word annotations for optical character recognition.

3 Detailed Explanation and Experimental Result

3.1 Hough Transform Method

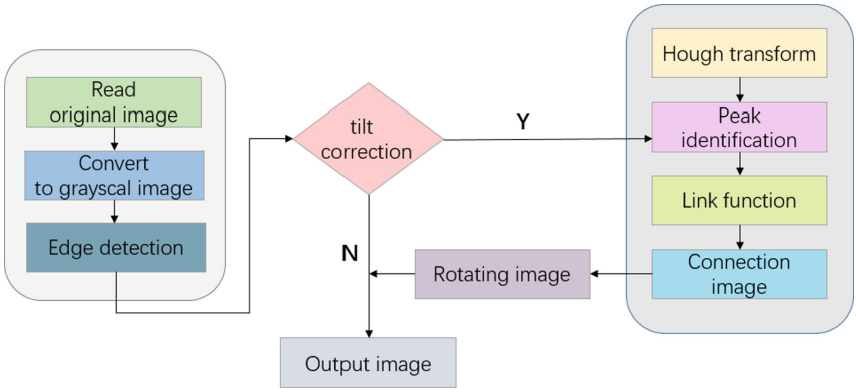


Fig. 6. Flowchart for preprocessing the invoice image by Hough transform.

Hough transform method edge connection flow chart is shown in Fig. 6. The flow is to read the original image and convert it into a grayscale image. Edge detection is performed by canny edge detection algorithm to obtain *binarized* edge images. Do a Hough transform on this edge image. Use the function *houghpeaks* for peak detection. The algorithm of the function *houghpeaks* is as follows: find the Hough transform unit containing the maximum value and record its position; set the Hough transform unit to 0 in the field of the maximum point found at the previous step; repeat this step until it finds the required up to the number of peaks, or until a specified threshold is reached. Once a set of candidate peaks has been identified in the Hough transform, it is left to determine if there are line segments associated with these peaks and their start and end. For each peak, the first step is to find the location of each non-zero point in the image that affects the peak. To do this, write the function *houghpixls*. Using the function *houghlines* to achieve straight edge joins, the function of the function *houghlines* is as follows: Rotate the pixel positions by $90^o - \theta$ so that they are roughly on a vertical line; sort the pixel positions by the rotated x value; use the function diff to find the split. Ignore the small split, which will merge adjacent segments separated by small blanks; return information for segments longer than the minimum threshold.

Fig. 7. Pre-processing process for invoice images in the experiment. a is the original invoice image; b is to convert the original image into a grayscale image; c is to convert gray image banalizations into edge image; d is the image obtained by the Hough transform and rotated.

Pre-processing process for invoice images in the experiment is shown in Fig. 7. The image has a certain angle of inclination and the background is striped. Since the features of the invoice are horizontal lines and vertical lines, the background also has horizontal lines, so it causes a lot of interference in the intelligent positioning of images. The principle of the Hough transform detection line will be based on the voting technique, and the straight line on the invoice with a large number of votes can be used to correct the tilt and solve the interference caused by the background.

3.2 Yolov3 Target Detection Network

The process of YOLOv3 training data is as follows: collect a large number of invoice images taken by mobile terminals from different angles, these images can contain different backgrounds; manually use the LabelImg marking tool to mark the images obtained in the previous step to generate a file of the format xml; The script file converts the xml file to a txt file. A training set is created for the Darknet-53 feature extraction network [12]; the training data is started, where the weight file will appear and the best weights will be imported into the test script; the intelligently positioned invoice image will be output.

YOLOv3 used logistic regression when predicting bounding box. Bounding box's coordinate prediction method:

$$b_x = \sigma(t_x) + c_x \tag{2}$$

$$b_y = \sigma(t_y) + c_y \tag{3}$$

$$b_w = p_w e^{t_w} \tag{4}$$

$$b_h = p_h e^{t_h} \tag{5}$$

where t_x, t_y, t_w, and t_h are the predicted outputs of the model. c_x and c_y represent the coordinates of the grid cell. The coordinates of the grid cell in the first column of the 0th row, c_x is 0, and c_y is 1. p_w and p_h represent the size of the predicted bounding box. b_x, b_y, b_w and b_h are the coordinates and size of the center of the predicted bounding box. $\sigma(t_x)$ and $\sigma(t_y)$ are the loss of coordinates using the squared error loss.

Fig. 8. YOLOv3 experimental results of intelligent positioning of invoice images.

Confidence reflects the accuracy of the bounding box which contains the object. The results of intelligent positioning of the invoice image captured by the smart phone are shown in Fig. 8. The IoU is a measure of the accuracy of detecting a corresponding object in a particular data set. The accuracy of intelligent positioning can be seen from

the IoU. The mean average precision can reach 92.5% when the IoU is set at 0.5. The mean average precision can reach 80.74% when the IoU is set at 0.7.

3.3 Optical Character Recognition

When we need to conduct text recognition on the invoice image that has been intelligently located and extracted, the following problems occur. Because the angle and height of the shooting will cause the character size on the image to be very different, and secondly, when the invoice is placed in different scenes, it also causes significant difficulties in identification. To solve the above problems, we use weak supervision. The learning framework performs precise character recognition on the extracted areas. The results identified by the weakly supervised learning framework are shown in Fig. 9. The accuracy of recognizing texts can reach up to 97.5%.

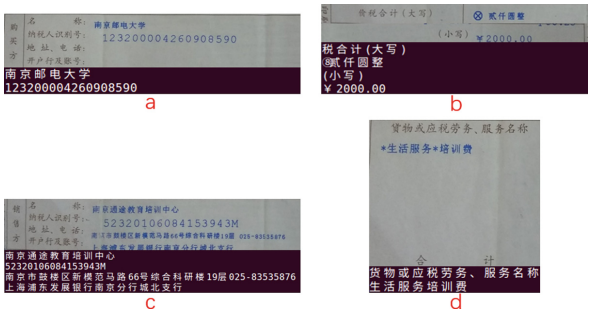


Fig. 9. The experimental results identified by the weakly supervised learning framework.

4 Conclusion

This paper has proposed a smart phone aided intelligent algorithm for invoice reimbursement system which can realize intelligent positioning and content recognition of invoice images. We applied the Hough transform detection linear principle to the preprocessing of invoice images. YOLOv3 target detection network was adopted to intelligently locate and extract invoice images taken in natural scenes. We used the weak supervised learning framework to perform accurate character recognition on the extracted invoice image. The accuracy of recognizing texts can reach up to 97.5%.

References

1. Bayar, S.: Performance analysis of e-Archive invoice processing on different embedded platforms. In: IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–4 (2016)
2. Sun, Y., Mao, X., Hong, S., Xu, W., Gui, G.: Template matching-based method for intelligent invoice information identification. IEEE Access **7**, 28392–28401 (2019)

3. Noble, F.K.: Comparison of OpenCV's feature detectors and feature matchers. In: International Conference on Mechatronics and Machine Vision in Practice (M2VIP), pp. 1–6 (2016)
4. Jiang, Y., Dong, H., El Saddik, A.: Baidu Meizu deep learning competition: arithmetic operation recognition using end-to-end learning OCR technologies. IEEE Access **6**, 60128–60136 (2018)
5. Pan, J., Yin, Y., Xiong, J., Luo, W., Gui, G., Sari, H.: Deep learning-based unmanned surveillance systems for observing water levels. IEEE Access **6**, 73561–73571 (2018)
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
7. J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *ArXiv*, vol. 2017–Janua, 2017
8. T. Zhi, W. Huang, H. Tong, H. Pan, and Q. B. T.-E. C. on C. V. Yu, "Detecting Text in Natural Image with Connectionist Text Proposal Network," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 56–72
9. B. Shi, X. Bai, and S. Belongie, "Detecting Oriented Text in Natural Images by Linking Segments," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3482–3490
10. X. Zhou *et al.*, "EAST: An Efficient and Accurate Scene Text Detector," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2642–2651
11. H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, and E. Ding, "WordSup: Exploiting Word Annotations for Character Based Text Detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4950–4959
12. Zhao, Y., Chen, Q., Cao, W., Yang, J., Gui, G.: Deep Learning for Risk Detection and Trajectory Tracking at Construction Sites. IEEE Access **7**, 30905–30912 (2019)