2016 International Conference on System Science and Engineering (ICSSE)
National Chi Nan University, Taiwan, July 7-9, 2016

An Intelligent Self-Checkout System for Smart Retail

Bing-Fei Wu, Wan-Ju Tseng, Yung-Shin Chen, Shih-Jhe Yao, Po-Ju Chang

{bwu, raccoon8215, cde566, a0991267281, bruce000011}@cssp.cn.nctu.edu.tw

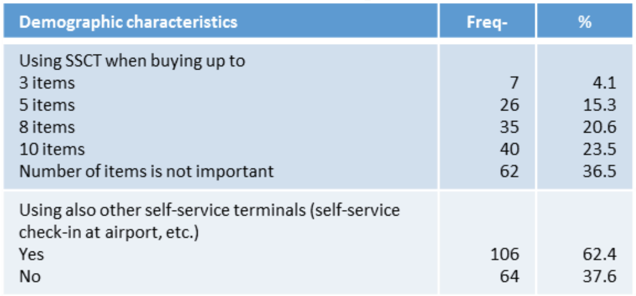
Abstract- Most of current self-checkout systems rely on barcodes, RFID tags, or QR codes attached on items to distinguish products. This paper proposes an Intelligent Self-Checkout System (ISCOS) embedded with a single camera to detect multiple products without any labels in real-time performance. In addition, deep learning skill is applied to implement product detection, and data mining techniques construct the image database employed as training dataset. Product information gathered from a number of markets in Taiwan is utilized to make recommendation to customers. The bounding boxes are annotated by background subtraction with a fixed camera to avoid time-consuming process for each image. The contribution of this work is to combine deep learning and data mining approaches to real-time multi-object detection in image-based checkout system.

Key Words: Self-checkout, Smart retail, Multi-object detection, Data mining, Deep learning

I. INTRODUCTION

The global self-checkout (SCO) shipments steadily grow in 2010 to 2019, according to Retail Banking Research (RBR) [1], which is a consulting firm provides services to organizations active in retail automation, banking and payments. In 2014, Sumak set a self-checkout system in the physical store in Slovenia for checkout effects [2]. Table 1 is the statistical results of customer acceptance to SCOs by using barcode scanning to make purchases. Consumers prefer SCOs to current checkout while buying more than five items and 62.4% of them had used other self-service terminals.

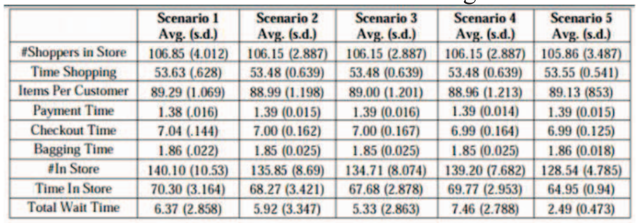
Table 1. Customer Acceptance to SCOs



Demographic characteristics	Freq-	%
Using SSCT when buying up to		
3 items	7	4.1
5 items	26	15.3
8 items	35	20.6
10 items	40	23.5
Number of items is not important	62	36.5
Using also other self-service terminals (self-service check-in at airport, etc.)		
Yes	106	62.4
No	64	37.6

Rossetti proposed the simulation modeling of customer checkout configurations and recorded its impact of checkout efficiency [3]. Checkout cases are divided into five scenarios. Scenario 1 is the traditional checkout system served by cashier scanning barcodes. Scenario 3 presents that customers choose SCO lanes by number of items. As a result, SCOs save one minute of total average waiting time for users compared to Scenario 1, see Table 2. Rather than dealing with the long lines, people can quickly pay for their purchases by SCOs especially during peak sales times.

Table 2. Customers' Total Waiting Time



	Scenario 1 Avg. (s.d.)	Scenario 2 Avg. (s.d.)	Scenario 3 Avg. (s.d.)	Scenario 4 Avg. (s.d.)	Scenario 5 Avg. (s.d.)
#Shoppers in Store	106.85 (4.012)	106.15 (2.887)	106.15 (2.887)	106.15 (2.887)	105.86 (3.487)
Time Shopping	53.63 (.628)	53.48 (0.639)	53.48 (0.639)	53.48 (0.639)	53.55 (0.541)
Items Per Customer	89.29 (1.069)	88.99 (1.198)	89.00 (1.201)	88.96 (1.213)	89.13 (853)
Payment Time	1.38 (.016)	1.39 (0.015)	1.39 (0.016)	1.39 (0.014)	1.39 (0.015)
Checkout Time	7.04 (.144)	7.00 (0.162)	7.00 (0.167)	6.99 (0.164)	6.99 (0.125)
Bagging Time	1.86 (.022)	1.85 (0.025)	1.85 (0.025)	1.85 (0.025)	1.86 (0.018)
#In Store	140.10 (10.53)	135.85 (8.69)	134.71 (8.074)	139.20 (7.682)	128.54 (4.785)
Time In Store	70.30 (3.164)	68.27 (3.421)	67.68 (2.878)	69.77 (2.953)	64.95 (0.94)
Total Wait Time	6.37 (2.858)	5.92 (3.347)	5.33 (2.863)	7.46 (2.788)	2.49 (0.473)

In this paper, an Intelligent Self-Checkout System (ISCOS) is proposed to address the non-barcode and camera reduction solution to current image-based checkout system. The system sums up the products overlooked by single camera set above the counter. Also, product detection prevents from fraud that users scan the cheaper items' labels rather than the more expensive ones [4]. We construct market database as training set to facilitate the research of shape matching and product recognition with emerging deep learning techniques, You Only Look Once (YOLO) [5] and ImageNet-trained CaffeNet[1] [6][7]. Each product is classified into one of shape categories by YOLO, using convolutional features to detect potential bounding boxes. After system has been processed by the training set, it deals with multiple instances of the same product and those with different brands in the image. Our work is also interactive that gives consumers shopping recommendations by comparing cross-market discounts in Taiwan.

II. RELATED WORK

There are several instances of machine vision checkout system for retail. Kroger was the first one to install a fully-automated scanner tunnels (FAST) in July 2010 [8]. The system is equipped with multiple image scanners that not only identify barcodes, but also analyze shapes, sizes and colors of package by using image recognition. Non-barcode items such as fruit and vegetable, need to be labelled in self-service before going through the conveyer belt. In February 2014, ECRS launched accelerated checkout system, called RAPTOR [9], providing customers a fast and convenient checkout transaction by capturing barcodes and images with over a dozen cameras. Nevertheless, items have to be spaced apart one after another onto the belt while using FAST and RAPTOR.

Recently, deep learning has been at the core of computer vision solutions to image classification [10], detection and localization. Approaches like R-CNN [11] and its variants[12][13], which achieve successful results on a variety of object detection tasks [14][15][16], use region proposals instead of sliding windows[17] to generate potential bounding

[1] @article{jia2014caffe,
Author = {Jia, Yangqing and Shelhamer, Evan and Donahue, Jeff and Karayev, Sergey and Long, Jonathan and Girshick, Ross and Guadarrama, Sergio and Darrell, Trevor},
Journal = {arXiv preprint arXiv:1408.5093},
Title = {Caffe: Convolutional Architecture for Fast Feature Embedding},
Year = {2014}
}

978-1-4673-8966-2/16/$31.00 ©2016 IEEE

boxes and identify objects by running classifiers on boxes. You only look once (YOLO), a unified object detector, encodes contextual information of entire image and deals detection work with single pipeline, from image pixels to class probabilities and predicted coordinates. Some machine cashiers for automatic bill calculation use these convolutional neural networks towards real-time performance to detect and classify objects simultaneously. Through the swift eye of computer vision, checkout system may compute retail product prices instead of RFID and barcode scanning.

III. DATA MINING

Items' information and images are downloaded for comparing prices and providing training data from the three markets' websites and three image search engines. There are 17 shape categories and 24 product classes in database respectively, which are constructed by automatic annotation avoiding the inconvenient issue of manual operation. Those images are employed as deep learning materials to complete multi-object detection task.

A. Data Collection

Products' data are gathered by ISCOS from three physical markets: Carrefour, RT-Mart, and A-Mart. Web crawler has the ability to scan large scale of product information and uses those controlled terms to index large document depositories effectively by classification, clustering, prediction, and regression [18]. The first task of database is to create a list of shape categories and product classes. After building the list, data mining controls commands to the downloader, and crawls names, prices, and image links based on commodity classes on three market websites, as shown in Fig. 1. Finally, commodity information and images are stored in the database in accordance with the corresponding categories.

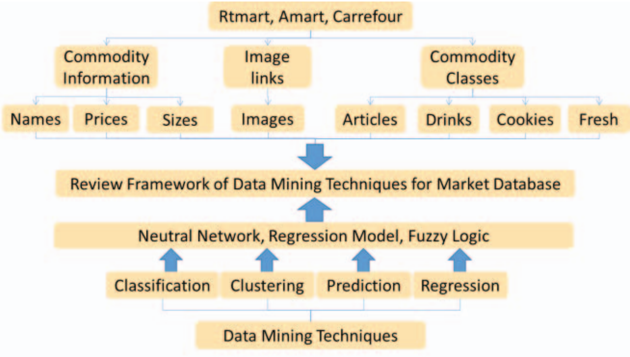


Fig. 1. Web Crawler Block Diagram

B. Automatic Annotation

In this work, product images including shape and logo content are gathered from Baidu, Yahoo, and Google. One of the most time-consuming processes is to determine objects and to annotate bounding boxes for each image. We introduce the automatic annotation method which finds out the moving objects by background subtraction with a fixed camera. It marks the location of objects' shapes by long-range shooting and extracts the product's bounding boxes by the near-range screening. Each shape category records up to 3,500 bounding boxes and each product class has about 10,000 images at any angle. Image collection is to ensure that images have at least a minimum number of objects for deep learning research.

C. Product Database

Classification and association models of data mining are used to classify products into categories in accordance with nature of items. Fig. 2 is a schematic view of database. After indexer crawls entire webpage content, each product is filtered by regular expression to construct database composed of names, prices, logos, and shapes, etc.

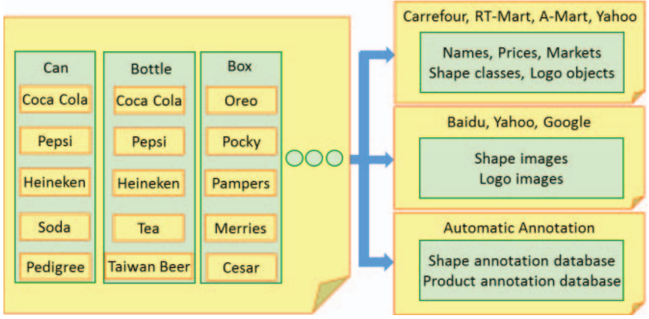


Fig. 2. Database Schematic

Table 3 gives the summary of databases crawled from websites. The number of logo classes, shape categories, and total number of images are different for each database.

Table 3. Database Summary

Dataset	Shape Database	Product Database	Shape Images	Logo Images
#Logo Classes	10	10	6	6
#Shape Classes	17	28	5	1
#Images	63,271	317,593	2,888	3,798
Image Resolution	640×480	456×417	541×517	776×718

D. Price Comparison and Shopping Suggestions

ISCOS compares prices with those in different markets and recommends consumers more preferential or more popular items, depending on sales volume extracted from inventory record. In addition, sales volume of each item is measured while customers pay for their purchases by ISCOS..

IV. PRODUCT DETECTION

Due to serving as checkout kiosk, the system is designed to achieve real-time performance. We decompose system into shape matching and product recognition to tackle multi-object detection problem as depicted in Fig. 3.

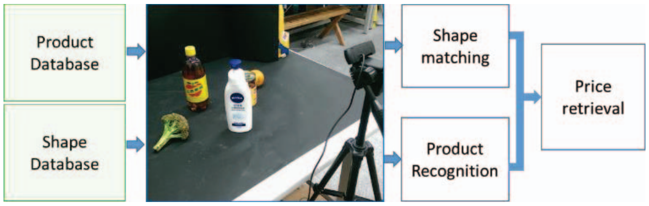


Fig. 3. System Diagram

We introduce YOLO as shape detector, locating products in each full frame and classifying them into shape categories such as bottle, can, circle, and box, etc. Next, we apply fine-tuned CaffeNet model, which is a variation on the well-known AlexNet architecture as the second stage to identify what products are present [10].

A. Shape Matching

An input image is divided into evenly spaced grids, that predict shape confidence scores and bounding boxes by single convolution network in real-time. A network architecture inspired by GoogLeNet model is adopted for shape matching as the first stage [19]. Furthermore, the output is adjusted to 9×9×27 predictions because products may be small in whole image and 17 shapes are defined. However, each grid cell only interprets two boxes with the same class. Under the above spatial constraint, products can overlap slightly but logos should present clearly in the frame.

B. Product Recognition

Except for R-CNN, most of the deep learning structure handle with the problem of single object classification. As we cascade YOLO detector, which is trained by automatic annotations, to scan the number of items in the frame, product recognition can be applied on every single cropped bounding boxes and system predicts bounding boxes and classify products' packages and brands. We started with the CaffeNet model developed by Berkeley Vision and Learning Center (BVLC). The network was initialized with ImageNet weights and only final fully connected layer (fc8) was fine-tuned.

V. Experimental Results

In this paper, we evaluate multiple object detection by a fixed overlooked camera. In order to solve the defect that most of the deep learning methods except for R-CNN are not capable of dealing with multiple items' identifications, this paper decomposes problem into two steps: shape matching and product recognition. The camera automatically locates objects' positions and identifies their shapes. Then, the system outputs classification results of each cropped single object image.

A. Training Parameters

The total training of YOLO over 30,000 iterations on a GTX970 GPU is set default configuration. Parameter of side is set at 7 in order to adapt smaller items present in the scene. For CaffeNet, the initial learning rate of 0.001 is lower as not to disrupt the ImageNet weights and dropped by 0.5 every 10,000 iterations. The network was fine-tuned over 100,000 iterations, momentum of 0.9, and batch size of 128.

B. Experimental Environment

Experimental environment is made with a single camera, which has an angle of depression of 33 degrees. The distance is measured at 36 centimeters between the camera and the platform. The installation is illustrated in Fig. 4.

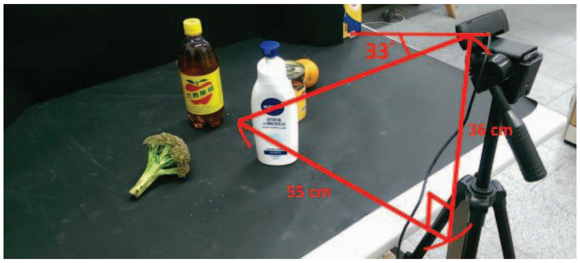


Fig. 4. Experimental Environment

C. Results

In our experiments, multiple products' positions are predicted by shape matching method. Fig. 5 and Fig. 6 display the detection result of one product and two items in the image. Two instances are chosen at random and are placed on the platform with slightly overlapping. However, the bounding boxes are still found clearly and quickly, and ISCOS searches the corresponding names and prices in the database based on the recognition result. Finally, the system prints out the total bill. The average detection time are 64.09 ms and 69.55 ms per frame of both one and two product detection.

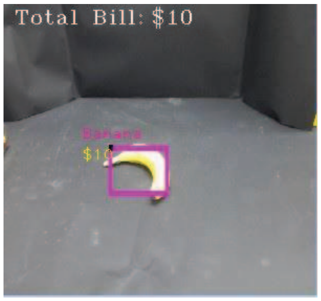


Fig. 5. Product Detection of One Item

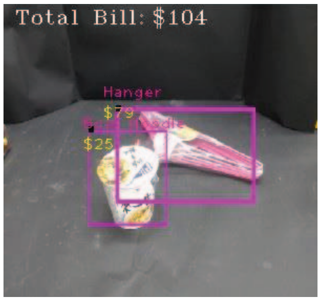


Fig. 6. Product Detection of Two Items

Fig. 7 provides the shape detection experiments of three items with logo and non-logo instances in the image. Even if items are put on the platform with slightly overlapping, ISCOS can also label these three bounding boxes stably. Total bill is calculated by the prices of three commodity and the result is displayed on the top-left of the screen. The detection time is 69.34 ms per frame.

The experiment of four products with overlapping to the utmost extent spends 75.57 ms to complete multiple object detection, see in Fig. 8. ISCOS labels multiple bounding boxes stably because the image frames still have some important features of products for detection. Commodity information and the total price simultaneously are presented to the user.

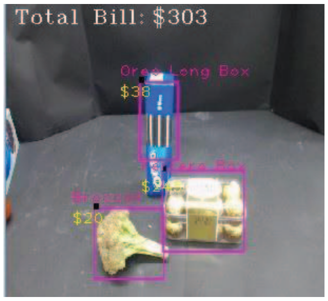


Fig. 7. Products Detection of Three Items

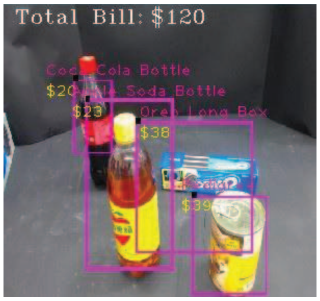


Fig. 8. Products Detection of Four Items

After the shape detection, we evaluate fine-tuned CaffeNet on 24 classes such as fruit, beverages, cookies, containers, and tools, etc. Each item presents in 1,000 frames at different scales and angles, as shown Table 4.

Table 4. Product Detection Accuracy

Name	Accuracy (%)	Name	Accuracy (%)
Oreo Long Box	80.7	Broccoli	72.4
Apple	97.1	Orange	100
Banana	58.0	Oreo Box	73.2
Bowl	98.5	Pedigree	78.2
Hanger	99.3	Pepsi Bottle	75.7
Coca Cola Bottle	93.9	Pepsi Can	70.1
Coca Cola Can	94.1	Pocky	76.5
Controller	36.4	Pliers	47.8
Lotion	91.2	Scissors	44.9
Mug	87.8	Apple Soda Bottle	95.6
Facial Cleanser	70.0	Apple Soda Can	49.1
Ferrero Box	65.7	Tea Bottle	77.9
Average Accuracy	76.48		

There are 750 times of experiments to simulate the consumer laying products on checkout conveyer belt at one time. Random numbers and categories of items are placed on the platform. Table 5 is the overall system performance analysis that gives the total time of detection including shape matching and product recognition. Error rate of the shape matching is calculated as :

$$Error\ Rate = \frac{|Box_{actual} - \ Box_{detected}|}{Box_{actual}} \times 100\%$$

, where Box_{actual} and $Box_{detected}$ are the number of actual and detected boxes in the presence of checkout platform respectively. Based on the above information and test results, we obtain the total accuracy as below.

Table 5. Overall System Performance Analysis

Numbers of Items Items	One Item	Two Items	Three Items
Total Times (ms)	64.09	69.55	69.34
Average Number of Bounding Boxes	1.016	0.962	0.857
Error Rate (%)	16.8	16.2	17.2
Total Accuracy (%)	66.4	65.7	64.1

D. *Performance Analysis*

As CaffeNet can only deal with single object classification, we cascade the detector and the localizer which gives coordinates to items appear in every frame [21]. Each input image is passing through a single neural network that predicts bounding boxes with average time of 69.6375 ms. The more items put on the checkout platform, the slower the processing is. However, machine vision checkout system introduces a time-saving solution for customers compared to barcode scanning purchases.

VI. CONCLUSIONS

The paper presents ISCOS, which contains product database and multi-object detection. We collect product data from three markets' web pages for price comparison, and gather training images from three image search engines. In order to reduce the time of manual annotation, the background subtraction technique is applied to bound products' locations automatically. Our work proposes the solution to implement SCO system, which identifies multiple products placed on the counter with slight overlapping and summing up prices by deep learning in real-time performance. SCOs change the checkout habits of the past, and ISCOS embedded with the function of multi-object detection and informative database increases the checkout convenience and efficiency.

On ISCOS, items are detected stably while there are less than three objects appear in the scene. Further, products may overlap slightly during image-based purchases. YOLO learns contextual information that results in model struggling with small objects appearing in the scene with other bigger objects. In addition, its coarse feature only handles with shape detector to facilitate product recognition. In this paper, ISCOS localizes products' positions even in the overlapping condition. It still relies on robuster, faster, and more stable object detection to make consumers use more intuitively.

ACKNOWLEDGEMENT

This work was supported by Ministry of Science and Technology under Grand no. MOST105-2622-E-009-013-CC2.

REFERENCES

[1] Retail Banking Research (2014, August 7). RBR_SCO_Press Release_070814. Message posted to http://goo.gl/tBKTCU

[2] Sumak, Bostjan, Maja Pusnik, and Marjan Heričko. "An empirical study of factors affecting the adoption of self-service checkout termi-nals in Slovenia."Information and Communication Technology, Elec-tronics and Microelectronics (MIPRO), 2014 37th International Conven-tion on. IEEE, 2014.

[3] Rossetti, Manuel D., and Anh T. Pham. "Simulation modeling of customer checkout configurations."Proceedings of the 2015 Winter Sim-ulation Conference. IEEE Press, 2015.

[4] Bobbit, Russell, et al. "Visual item verification for fraud prevention in retail self-checkout." Applications of Computer Vision (WACV), 2011 IEEE Workshop on. IEEE, 2011.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640 (2015).

[6] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3 (2015): 211-252.

[7] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast fea-ture embedding."Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[8]Planet Retail (2014, May 21). Insight: FAST not furious - is this the end for the big basket problem? Message posted to http://goo.gl/UNf8Ea

[9] Retail & Loyalty NEWS (2014, January 17). With the Introduction of RAPTOR, ECRS Looks to Change How Groceries Are Purchased. Message posted to http://goo.gl/g8Bq1N

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

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Con-ference on, pages 580–587. IEEE, 2014.

[12] R. Girshick. Fast R-CNN. In Computer Vision, 2015 IEEE International Conference on, pages 1440-1448. IEEE 2015.

[13] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015.

[14] Lienhart, Rainer, and Jochen Maydt. "An extended set of haar-like features for rapid object detection." Image Processing. 2002. Proceedings. 2002 International Conference on. Vol. 1. IEEE, 2002.

[15] Viola, Paul, and Michael Jones. "Robust real-time object detection." International Journal of Computer Vision 4 (2001).

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

[17] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010.

[18] Deepa, V. K., and J. Rexy R. Geetha. "Rapid development of applications in data mining." Green High Performance Computing (ICGHPC), 2013 IEEE International Conference on. IEEE, 2013.

[19] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.