IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON GREEN INTERNET OF THINGS

Received June 17, 2020, accepted July 16, 2020, date of publication August 3, 2020, date of current version August 21, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3014047*

# Design of Smart Unstaffed Retail Shop Based on IoT and Artificial Intelligence

**JIANQIANG XU[1], (Member, IEEE), ZHUJIAO HU[2], ZHUO ZOU[2], (Member, IEEE), JUNZHONG ZOU[1], XIAOMING HU[3], (Senior Member, IEEE), LIZHENG LIU[2,3], (Member, IEEE), AND LIRONG ZHENG[2], (Senior Member, IEEE)**

[1]East China University of Science and Technology, Shanghai 200237, China
[2]School of Information Science and Technology, Fudan University, Shanghai 200433, China
[3]Optimization and System Theory Department of Mathematics, Royal Institute of Technology (KTH), 11428 Stockholm, Sweden

Corresponding author: Lizheng Liu (lzliu@fudan.edu.cn)

This work was supported in part by the NSFC under Grant 61876039 and Grant 62011530132, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2018SHZDZX01, in part by the ZJ Lab, in part by the China Postdoctoral Science Foundation under Grant 2020M670993, and in part by the Shanghai Platform for Neuromorphic and AI Chip (NeuHeilium).

**ABSTRACT** Unstaffed retail shops have emerged recently and been noticeably changing our shopping styles. In terms of these shops, the design of vending machine is critical to user shopping experience. The conventional design typically uses weighing sensors incapable of sensing what the customer is taking. In the present study, a smart unstaffed retail shop scheme is proposed based on artificial intelligence and the internet of things, as an attempt to enhance the user shopping experience remarkably. To analyze multiple target features of commodities, the SSD (300 × 300) algorithm is employed; the recognition accuracy is further enhanced by adding sub-prediction structure. Using the data set of 18, 000 images in different practical scenarios containing 20 different type of stock keeping units, the comparison experimental results reveal that the proposed SSD (300 × 300) model outperforms than the original SSD (300 × 300) in goods detection, the mean average precision of the developed method reaches 96.1% on the test dataset, revealing that the system can make up for the deficiency of conventional unmanned container. The practical test shows that the system can meet the requirements of new retail, which greatly increases the customer flow and transaction volume.

**INDEX TERMS** Unstaffed retail, Internet of Things, vending machine, artificial intelligence, SSD.

## I. INTRODUCTION

Nowadays, a growing number of businesses and consumers are stressing the efficiency and experience of consuming shopping. As fueled by the progress of the Internet of Things (IoT) and Artificial Intelligence (AI), and the spread of smartphones and mobile payments, the unstaffed retail shopping is being popularized. According to the survey data from iiMedia Research, it is estimated that by 2022, 245 million consumers will shop through unmanned retail shops in China, and the transactions volume of unmanned retail will exceed 1.8 trillion CNY. IoT technology allows for connecting sensors and intelligent hardware to the internet, it is a growing ubiquitous concept that has impacted considerable aspects of human life [1]–[3]. IoT-based service has been applied in the novel retail shopping fields, thereby offering a set of interactive shopping solutions [4], [5]. The customers can complete the whole shopping process by self-service, and the whole consumption process appears to be naturally smooth.

In the unstaffed retail shop, the smart vending machine acts as a crucial instrument. The conventional unmanned container faces high commodity loss rate, and it is not convenient to carry out goods management. Intelligent container that adopts RFID technology is an effective solution [6], whereas the late operation costs are too high. Some intelligent containers using gravity sensing method are capable of detecting the commodity being moved, whereas it cannot sense what the customer is taking. Emerging interest has occurred in integrating machine vision technology into unmanned retail, including Amazon, Alibaba, etc. Amazon presented its automated Go grocery store concept based on machine vision in 2016 [7], which can record the items that the consumers

The associate editor coordinating the review of this manuscript and approving it for publication was Jun Wu.

pick up. Alibaba has rolled out an experimental cashier-less store named ''Tao Cafe'' [8]. In such stores, customers will automatically make a purchase via their smartphones without the need to head to a register and be able to leave the store with their items in hand. Megvii Technology proposes a large scale retail product checkout dataset to address the problem that the lack of a high-quality images dataset of the products to purchase [9].

Recently, deep learning has displayed its potential in the versatile and highly variable tasks of a range of fine-grained object detection and recognition. It is capable of recognizing patterns with extreme variability, and with robustness to distortions and simple geometric transformations. The algorithms have been applied in visual analysis (e.g., object detection, object recognition, and intelligent service robot) [10], [11]. Mainstream methods for object detection consist of:

1) Region-based convolutional neural network (R-CNN) model is capable of deriving Fast R-CNN and Faster R-CNN) [12], [13], combining region proposal and convolutional neural network (CNN) classification;
2) You only look once (YOLO) model [14];
3) Single Shot MultiBox Detector (SSD) model [15], integrating the anchor mechanism of Faster R-CNN and the regression idea of YOLO, which significantly enhances the accuracy of locating and classification;
4) MASK-RCNN [16].

As fueled by the advancement of large-scale GPU parallelization platform and other technologies, most deep neural networks are run based on centralized computing [17], [18]. The emerging interest has occurred in integrating AI technology into the retail industry, the AI method is applied for goods detection from different perspectives. [19] Proposes a scale-aware object proposal detection framework for crowded product detection in supermarket scenarios. [20] Proposes a single deep convolutional neural network termed as DiffNet to identify different objects in a pair of images taken from the same environment. A new scheme is proposed to enhance the operation speed of YOLOv3, which is used for fast detection of objects for vending machine. In [21], a fish-eye camera was used to obtain the images of goods, and MASK-RCNN algorithm was adopted to implement automatic recognition of items. However, the test scene is relatively simple as being consistent with the real application scene, and this study did not solve the recognition rate problem by the mutual occlusion of goods in the vending machine. The fisheye correction algorithm was developed to correct the distortion of image, whereas the computing speed decreased. In the present study, we further refined the smart unstaffed retail shop scheme. A method based on deep learning is developed to get the classification results by analyzing the top features of items and body features of the items; sub-prediction structure is designed in SSD algorithm, as an attempt to up-regulate the recognition rate of the items in the vending machine in the practical application scenarios. The dual camera system is adopted to capture the images of items and reduce the influence of image distortion. The data set is expanded to 18,000 pictures of unmanned containers in practical application scenarios.

The rest of the present study is organized as follows. Section II discusses smart unstaffed retail shop scheme. Section III introduces the proposed AI algorithm. Section IV reports the experiment results. Section V gives a conclusion.

## II. SMART UNSTAFFED RETAIL SHOP SCHEME

### A. THE ARCHITECTURE OF THE SMART UNSTAFFED RETAIL SHOP

Due to the saturation of online flow, as well as the rise of customer acquisition costs and labor costs, unmannered retail becomes a hotspot. The increasing retail market of consumption goods has further formed the basis of ''novel retail'' style. As one of the novel retail style, the smart unstaffed retail shop should solve the following two main problems:

a) Consumer identification: Identification based on smartphone information binding and identification based on biometric features recognition can be exploited for consumer identity authentication. The former method is generally implemented by SMS payment and scanning code payment. The latter complies with facial recognition and palm recognition methods to complete the payment. The consumers are not required to use a mobile phone, and the whole process is senseless.
b) Commodity recognition: The mainstream commodity identification consists of gravity sensing, RFID identification and machine visual recognition. In theory, commodity recognition based on machine vision can be exploited for quantity statistics and object recognition simultaneously, allowing the consumers to gain better shopping experience.

The logical architecture of the proposed smart unstaffed retail shop is illustrated in Figure. 1. It is split into four layers, namely, the front-end layer, the business management layer, the server & service layer and the database layer.

The front-end covers cameras, intelligent unmanned retail container, and IoT controller, directly interacting with consumers while harvesting and transferring data. For the large space in the vending machine, using monocular camera is easy to have incomplete images, and the camera with fisheye lens employed in such scenario will be subject to fisheye distortion [22]. To address the mentioned shortcomings, dual cameras are adopted here to capture images of the inner items. The vending machine refers to retail goods storage equipment for free purchase. These instruments are interconnected with servers through the IoT control unit.

The business management layer consists of enterprise resource planning (ERP) system, namely, a shopping management platform, on which user management, order management, stock keeping units (SKU) management, interface management and payment service are implemented. The SKU management includes specific information of the commodity (e.g., commodity type, name, specification, market

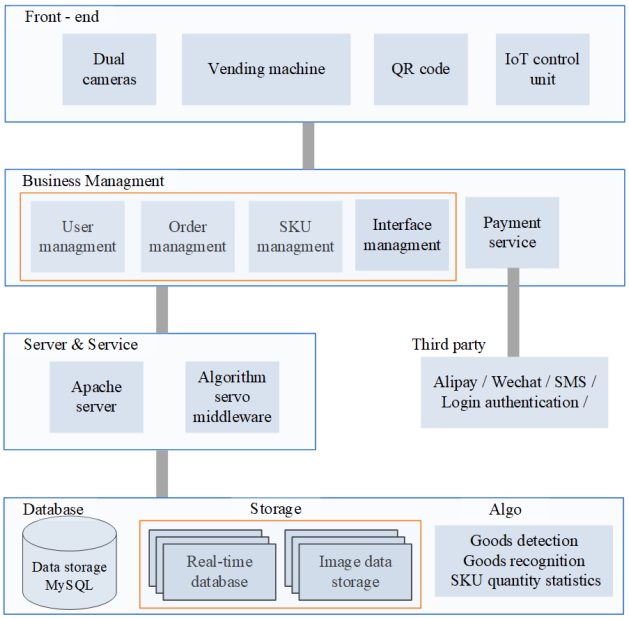



**FIGURE 1.** The logical architecture of the smart unstaffed retail shop.

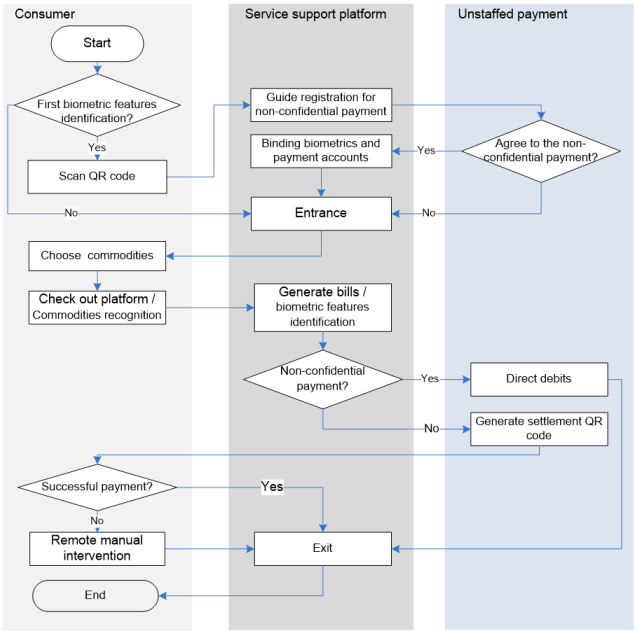



**FIGURE 2.** The entire service flow in the smart unstaffed retail shop.

price, selling price, detail chart, expiration date, and SKU). User management connects user account, provides account ID, key, distributes digital certificate, etc. Order management largely records users' order and transaction process. Third party service (e.g., WeChat and Ali-pay) is connected to the platform via this layer.

The server & service layer consists of the algorithm servo middleware and the Apache server. The Apache server offers the basic HTTP service. The algorithm servo middleware links the background program layer and the algorithms.

The database layer is primarily composed of MySQL database and file storage services. In such design, a distributed storage method is adopted to store basic data, image files, and the algorithms (e.g., goods detection, goods recognition, and SKU quantity).

### B. UNMANNED RETAIL SHOPPING PROCESS

The smart unstaffed retail shop is split into three parts, namely, consumer, service support platform and unstaffed payment platform, at the application logic level. Figure 2 presents the entire service flow in the smart unstaffed retail shop with non-confidential payment. The registered consumers enter the shop by biometric features identification, which is also applied in bills payment. If the consumer is not registered, he / she will be guided to register after scanning the dynamic QR code. The commodities that they select are recognized automatically through AI method, and the payment process covers bills generation that is automatically processed in the background. Biometric identification (e.g. facial recognition or fingerprint) can be exploited to identify whether the consumer is on the blacklist; this effort improves the security of the whole shopping. Using AI methods, consumers can achieve "Take and Go" in the process of unmanned retail shopping.

## III. PROPOSED AI METHOD

AI detection and recognition of SKU is critical to the smart unstaffed retail shop. In the popular unstaffed retail shop, the vending machine is largely used to store the items (e.g., drinks), most of which is bottled or canned. Here, we take these bottled and canned drinks as identified objects.

AI detection and recognition of SKU is a technology that locates the bounding-box of the goods from a raw inner container image captured by the binocular camera; it is critical to a smart vending system. The goods can be detected based on color feature, mathematical morphology, texture feature and depth feature. For the first three characteristics, the speed of CPU implementation can reach 40 FPS (seconds per frame), whereas the recalling rate and mean average precision will be practically down-regulated by the pattern noise, background interference and poor quality of the image. Moreover, for depth feature based object detection method, the three major approaches, namely, 1) Faster R-CNN, 2) YOLO and 3) SSD, can exhibit high accuracy rate. After these three models are trained on the union of VOC2007 and VOC2012 dataset, the comparison between SSD, Faster R-CNN, and YOLO on Pascal VOC 2007 test set shows SSD ($300\times300$) outperforms YOLO in detection speed, and it is close to Faster R-CNN in accuracy. Accordingly, the SSD ($300 \times 300$) is employed to extract the location information of the goods.

### A. IMPROVED SSD NETWORK

The SSD is primarily to extract different feature layers in basic network like visual geometry group network (VGG), and add the specific convolutional layer, which progressively reduces the size of feature graph; thus, the semantics can be improved to be more advanced. Each point in the feature

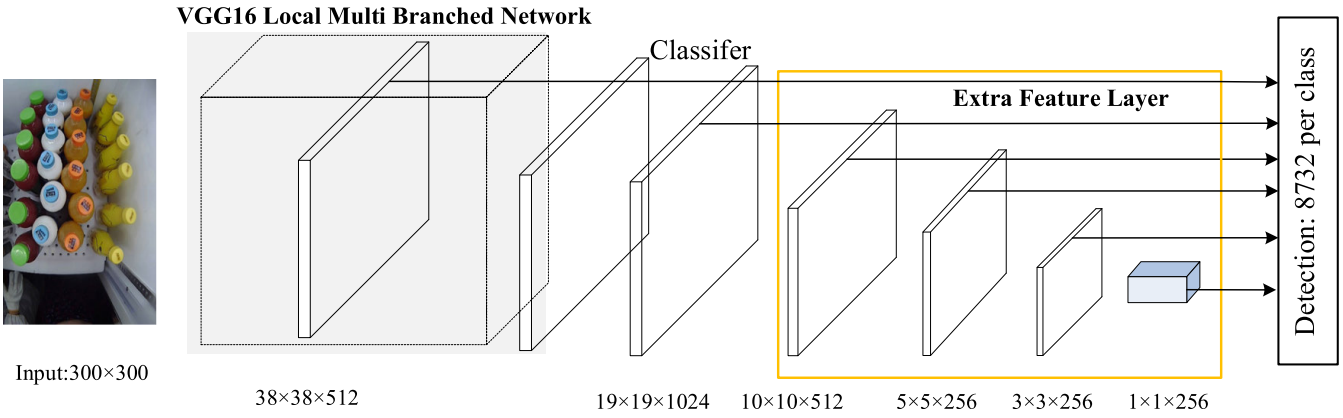



**FIGURE 3.** The diagram of the architecture of SSD (300 × 300) network for product detection.

layers corresponds to an anchor of different size that corresponding to different sensory fields in the original image. By the convolution of two 3 × 3 channels to predict the score and position of respective type, regression training is performed on it, and finally non-maximum inhibition is used to remove those redundant bounding box. In [16], the method of identifying objects in images with SSD is describe. Such approach can be applied to detect items in vending machine, whereas it cannot distinguish the different items that have the same shape. In view of this, we will start from the aspect of model adjustment and add color feature information with special weight to deal with object recognition with the identical shape.

The architecture of SSD (300 × 300) network for product detection is illustrated in Figure. 3, the VGG16 serves as the fundamental network for SSD, and the part in the yellow box is the feature extraction layer added on the basis of VGG16, each point of these feature layers corresponds to different size of anchors. The prediction in SSD will be carried out on the 5 feature maps selected in the previous steps besides object detection on the final feature map selected.

#### 1) SKIP LINKS OF CONVOLUTIONAL NEURAL NETWORKS

In the early image classification task, problems (e.g., training difficulties) emerge with the deepening of the network depth. The method that link different convolutional layers was proposed to learn different semantic combinations [23]. As neural networks get deeper, features tend to be more difficult to be learnt. By adding a shortcut path, the learning process can shift from directly learning features to adding features to previously learned features to acquire better features. Assuming that the final expectation is H(x), and each layer generates F(X) after activation function, the input x in the pre-layer and the output of current-layer are directly added to the next layer, namely, $H(x) = F(x) + x$. Another form of extension is to jump to multi-layer, such multi-layer jumping can combine multi-layer features to learn other meaningful data. Learning aims to go from learning complete information to learning residuals, so effective features can be more easily learnt.

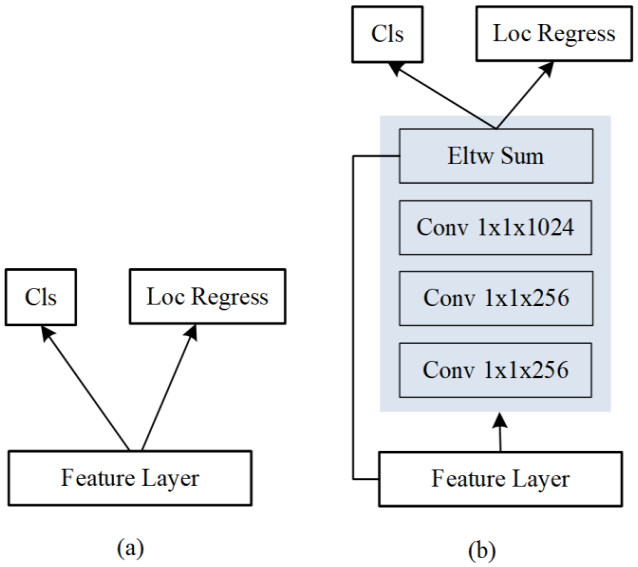



**FIGURE 4.** (a) The prediction module of the original SSD. (b) The residual block added a jumping link.

#### 2) PREDICT STRUCTURE

In the original SSD algorithm, objective function is directly applied in the specific layers of feature graph, and L2 regularization method is adopted. To enhance the accuracy of model, paper [24] suggested that the accuracy will be enhanced with the improving of the subnetwork of each task. Given this result, a residual block is added on each prediction layer, Figure.4 (a) gives the original model structure of SSD, and Figure.4 (b) illustrates the design that a jumping link is added to the residual block, a convolutional layer of 1 × 1 is added to the jumping link.

### *B. METHOD OF MODEL TRAINING*

#### 1) IMAGE PRE-PROCESSING AND PRE-TRAINING WEIGHTS

Multiple image preprocessing schemes were added to the original SSD model to improve the generalization ability of the model, e.g., image random clipping, image resize, image color distortion and so on. To increase the robustness of

the model to the size of goods images, data augmentation is performed on training imagines, which will significantly impact the network performance. For a sample fragment, setting the overlap between the clip and the target to 0.1, 0.3, 0.5, 0.7, and 0.9, respectively, and resize after clipping. To handle object with different sizes, we can achieve almost the same result using the feature map on different layers of the same network. This is because there are more details preserved in the lower feature map, which can be employed to improve the segmentation results. Accordingly, lower feature map is applied in the goods detection method. In this study, the VGG16 network acts as the main network but removes the full connection network. The original weights of VGG16 network are adopted to initialize the backbone network of SSD, and the rest network layers are randomly initialized. The scheme of network initialization determines the final training effect; the weight of pre-training is adopted to initialize the entire network model without any oscillation. Furthermore, a good convergence is conducted, and a good effect on the test set is achieved.

#### 2) DEFAULT BOX MATCHING STRATEGY

The SSD network structure generates several pre-defined candidate boxes commonly called default boxes to enhance efficiency [16], it is a rectangular box with different height to width ratios and different scales in a range of positions or layers on the feature map. Mapping these rectangular boxes to the original image will see the different receptive field size. During the training period, which default box correspond to ground truth detection should be ascertained. For respective ground truth box, the default box is selected from different positions, different aspect ratios, as well as different sizes. The matching rule is intersection over union (IOU) of ground truth box and default box, and the IOU threshold is set to 0.5. Such strategy allows us to get multiple default box with higher scores instead of just choosing one default box with the highest score. For the default boxes that do not match, the ratio of positive and negative example is set to 1:3 to reduce the imbalance of positive and negative example data, and the negative samples are sorted in line with confidence coefficient.

#### 3) DEFAULT BOX ASPECT RATIO SETTING

In the original SSD framework, the ratio of default box is set to 2 or 3. The clustering method based on K-means is adopted to obtain the ideal ratio in this study, the square of box area is taken as the feature of clustering. Two clusters are set at beginning, and we successively increase the number of clusters if the faults can improve 20%, seven clusters were collected, and the results are listed in Table 1.

Based on the experimental result above, we add more aspect ratios as 1.6, 2.0, 3.0 etc.

#### 4) MODEL PARAMETERS

In the present study, the input images are first resized to $300 \times 300$ and then inputted into the network model to obtain feature mappings of different sizes. The FC6 layer and FC7 layer of VGG16 are converted into a convolutional network; all the dropout layers and the FC8 layer are removed, pool5 is converted from $2 \times 2 - s2$ to $3 \times 3 - s1$. conv4_3, conv7, conv8_2, conv9_2, conv10_2, and conv11_2 are used to predict the location and category confidence, the feature map of these layers is extracted, and the multiple bound box of different sizes is constructed at each pixel on these feature map layers. Subsequently, detection and classification are performed separately to generate a bound box. The bound boxes obtained by different feature maps are combined; some overlapped or incorrect bound boxes are suppressed using non-maximum suppression (NMS) method to select the area with the maximal confidence value and the object to yield the final bound box set. Table 2 details some parameters of SSD ($300 \times 300$) algorithm of this study.

**TABLE 1.** Cluster error rate results of training data under different W/H ratios.

| W/H | 1.0 | 0.7 | 0.5 | 0.3 | 1.6 | 0.2 | 2.9 |
|---|---|---|---|---|---|---|---|
| Max(W/H,H/W) | 1.0 | 1.4 | 2 | 3.3 | 1.6 | 5 | 2.9 |
| Error rate (%) | 22.6 | 21.3 | 19 | 13.6 | 12.8 | 6.7 | 4.1 |

**TABLE 2.** Different feature layers are used for prediction.

| Feature layers | Conv4_3 | Conv7 | Conv8_2 |
|---|---|---|---|
| Size of feature layers | $38 \times 38$ | $19 \times 19$ | $10 \times 10$ |
| Anchor sizes | (21, 45) | (45, 99) | (99, 153) |
| Anchor ratios | 2, 0.5 | 2, 0,5, 3, 1/3 | 2, 0,5, 3, 1/3 |
| Prior box | 4 | 6 | 6 |
| **Feature layers** | **Conv9_2** | **Conv10_2** | **Conv11_2** |
| Size of feature layers | $5 \times 5$ | $3 \times 3$ | $1 \times 1$ |
| Anchor sizes | (153, 207) | (207, 261) | (261, 305) |
| Anchor ratios | 2, 0,5, 3, 1/3 | 2, 0.5 | 2, 0.5 |
| Prior box | 6 | 4 | 4 |

## IV. EXPERIMENTS AND RESULTS

### A. DATA SET PREPARETION

For the insufficient SKU standard data sets, considerable SKU images are captured by the camera in the vending machine. The graphical image annotation is time-consuming, the collected images are annotated bottle cap area and bottle body area, and the category of the area is labelled. Lastly, an XML file is generated to store the region and category information.

The wrong label positions are deleted, and the wrong tags are fixed by artificial selection. Moreover, the undetected target images are also added. The XML file is delved into programmatically to get the type and number of items in

each picture, since each picture is named strictly before image collection, some error information can be identified by comparing the type and name of the parsed items. After the data cleaning, the data is converted to be recognizable by the network, largely converted as LMDB. The finished data acts as the training sample data for the detection and classification. The graphical image annotation tool Labelimg is employed to generate data sets, and the segmentation region of commodities is described as accurately as possible with multiple key points.

In the mentioned work, a total of 18,000 images in different scenarios with 20 different types of SKU are collected. These images cover different SKU combinations and different placement; the distribution of the item types in the 18,000 images is listed in Table 3 where 1-20 represent these 20 different types of SKU. 90% of these data is applied in training, while 10% of them is in model testing. To ensure that there is no overlap between the data in the test data set and the training data set, the test data set is developed in a random arrangement.

**TABLE 3. Distribution of item types.**

| Type | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Numbers | 880 | 872 | 932 | 930 | 922 |
| Type | 6 | 7 | 8 | 9 | 10 |
| Numbers | 878 | 880 | 925 | 912 | 894 |
| Type | 11 | 12 | 13 | 14 | 15 |
| Numbers | 915 | 906 | 911 | 868 | 896 |
| Type | 16 | 17 | 18 | 19 | 20 |
| Numbers | 898 | 903 | 889 | 875 | 904 |

### B. MODEL TRAINING

The training system is assessed on PC with Intel(R) Xeon(R) CPU I7-7700 with 8 cores and two GTX 1080 GPUs of NVIDIA GeForce; each GPU covers 8GB memory and 9 TFLOPS processing performance. Tensor Flow [25] framework is adopted to assess the accuracy of the proposed model. Each batch is set as 32 pictures; the learning rate is set as 0.001 at the beginning and iterates 40k times. Subsequently, the learning rate decreases to 0.0001 and iterates to 60k. Lastly, the learning rate is set as 0.00001 and iterates to 70k. In this study, the transfer learning is exploited to accelerate the convergence of the model and narrow the training time [26]; feature vectors are extracted for transfer learning. First, the data of VOC07 trainval and VOC12 trainval are adopted to train the enhanced SSD network, the training set using VOC07 and VOC12 contains around 16,000 images. The major motivation is the similarity between our data set and VOC data set. The novel model is initialed by the weighs gained after training, the newly added predictive layer weights are randomly initialized, and the learning rate is set as 0.0001.

### C. INFERENCE AND COMPARISON EXPERIMENT

Based on the data set, 1800 samples are taken as the testing dataset. All item characters in testing dataset are labeled, and no overlap exists between the testing dataset and the training dataset. The test images originate from complex practical scenarios, which makes the recognition difficult due to the density. In the practical scenarios of vending machine, the identical items would be placed in a row, the results of SSD item recognition models are presented in Figure. 5.

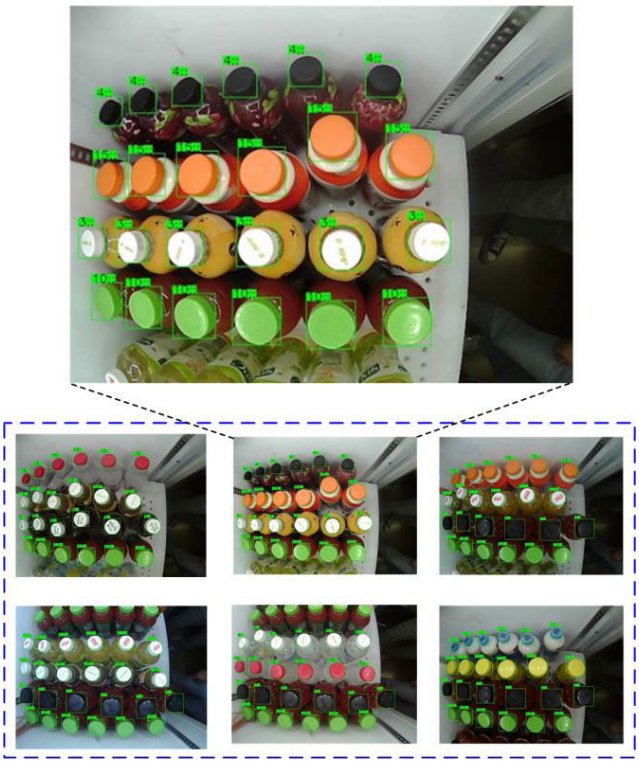


**FIGURE 5. Examples of goods detection results.**

### D. COMPARISON WITH ORIGINAL SSD

The comparison experiments are carried out between the original SSD (300 × 300) model and the proposed SSD (300 × 300) model, the main difference of the two models is that the sub-prediction structure is added to the later. We carry out controlled experiments to examine how each setting or component influences performance. The same computing system described above is employed for these two models, and they use the same training data set and inference data set. Figure. 6 illustrates the detection precision of these two models for 20 type of items. The mean average precision (mAP) of a range of models is listed in Table 4.

It can be seen from Table 4 that the mAP of the proposed SSD (300 × 300) is up to 96.1%, and the mAP of the original SSD (300 × 300) model is 91.64%. The detection time of the two models are 22ms and 21ms, respectively. The experimental results show that the proposed SSD (300 × 300) model has higher detection precision than the original


**FIGURE 6.** A comparison of precision of different models for 20 type of items between the proposed SSD and the original SSD.

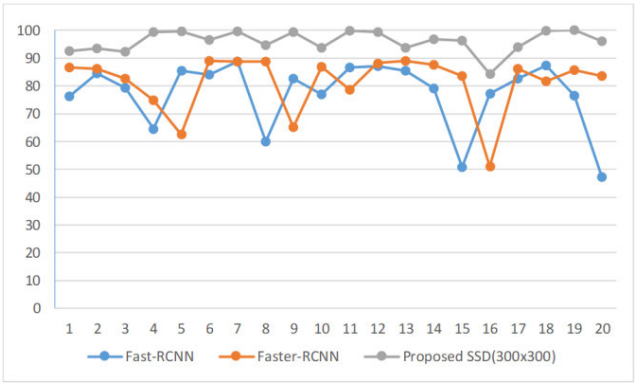



**FIGURE 7.** A comparison of precision of different models for 20 type of items.

**TABLE 4.** A comparison of mean average precision and detection time between the original SSD and the proposed SSD.

| Models | Original SSD (300×300) | **Proposed SSD (300×300)** |
|---|---|---|
| mAP % | 91.64 | **96.1** |
| Time (ms) | 21 | **22** |

**TABLE 5.** A comparison of mAP for different models.

| Models | Fast R-CNN | Faster R-CNN | **Proposed SSD (300×300)** |
|---|---|---|---|
| Network | VGG | VGG | **VGG** |
| mAP % | 77.1 | 81.5 | **96.1** |

SSD (300 × 300) model, the experimental results show the residual block with jump link can better detect small objects, and thus significantly improves the detection precision of the model. In the proposed SSD model, a convolutional processing is added to the sub-prediction structure, such addition increases the detection time by 1ms.

### E. COMPARISON WITH RCNN

During the experiments, the proposed two stage SSD (300 × 300) detection model are assessed and compared with the other two stage R-CNN detection models on the same data set, i.e. Fast R-CNN [13] and Faster R-CNN [14]. Region-based convolutional neural networks (RCNNs) is essentially a feature extractor. Fast RCNN does not put every region proposal into the extraction when extracting features, but extracts features of the whole map by means of coordinate mapping. Faster RCNN uses region proposal network to replace selective search, which can help improve the computing speed. The precision of different detection models for 20 type of items is illustrated in Figure. 7. It can be viewed that the precision of the proposed SSD (300 × 300) model is higher than other two models.

The mean average precision (mAP) of a range of models is listed in Table 5, the mAP of Fast R-CNN and Faster R-CNN 77.1% and 81.5%, respectively, it shows that the proposed SSD (300 × 300) model outperforms than other two models in detection. The main reason is that the multi-layer network features are utilized in SSD, it combines with the prediction results of multiple feature maps of different sizes, and objects of different sizes can be processed.

### F. CUSTOMER FLOW AND TRADING VOLUME TEST

20 smart vending machines in five convenience shops were installed to operate under the new retail mode, each store has four smart vending machines, aiming to test the impact of smart vending machines on customer flow and customer transactions. In order to avoid the problem of proximity interference of similar products, we put similar products at a distance. During the one-month test, the whole system worked well and there was no problem of product misidentification. We did the statistics of the customer flow and transaction volume of the five stores operating in the new retail mode for a month, the transaction volume is defined as the number of transactions here. Figure 8(a) shows the comparison of the customer flow of the five shops in the traditional mode and the new retail mode respectively in a month. The statistics in the traditional mode are from the data in the month before the installation of the smart vending machine. As can be seen from the figure, compared with the traditional retail model, the customer flow of each shop under the new retail model increased significantly, and the overall customer flow increased by 21.7%. Figure 8(b) is the comparison of the transaction volume of the five shops in a month under the traditional mode and the new retail mode respectively. The transaction volume also increased under the new retail mode, with the overall transaction volume increasing by 26.8%. Based on the analysis of the growth rate of customer flow and transaction volume, it can be seen that the transaction rate is also increasing, which indicates that the new retail mode adopting artificial intelligence technology can promote the development of offline commodity retail.

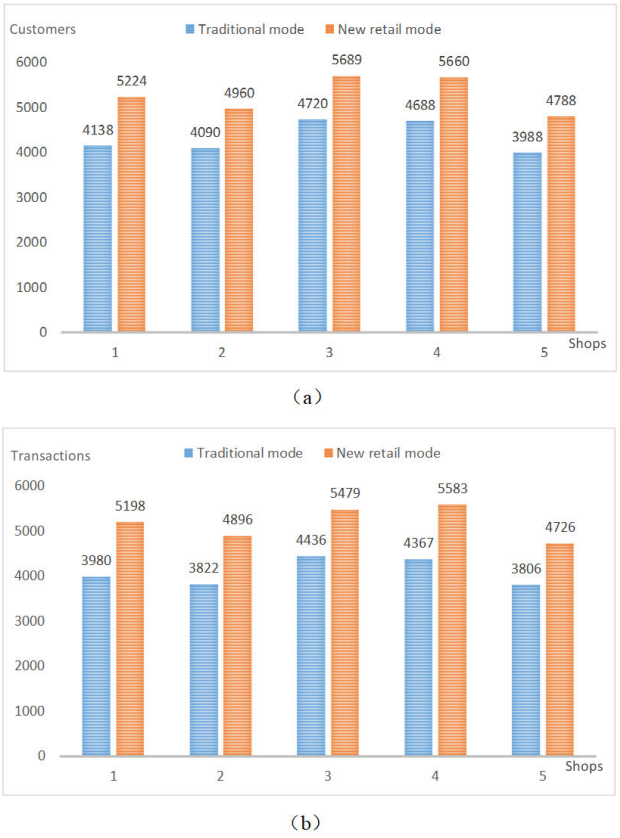



**FIGURE 8.** (a) A comparison of the customer flow of the five shops in different retail mode. (b) A comparison of the transaction volume of the five shops in different retail mode.

### G. DISCUSSION

#### 1) TRAINING CONVERGENCE

In this study, the experimental results indicate that the proposed SSD (300 × 300) model can more effectively solve the problem of commodity placement. In the preliminary experiment, however, it is reported that the SSD algorithm model based on the Tensorflow framework do not converge on the recognition of various goods. Accordingly, this problem is simplified, and the bottle cap part of the goods is made able to be well detected by distinguishing the foreground from the background. Furthermore, the position regression also achieves the desired effect.

The negative samples generated by default box after matching will lead to the dissatisfied convergence in the training. According to our training experience, if the proportion of negative sample and positive sample is too large (for example, greater than 10:1), the training process is difficult to converge or the classifier may tend to identify all samples as negative samples.

#### 2) FACTORS INFLUENCING THE RESULTS OF MODEL TRAINING

During the experiment, it is reported that the data set collection is incorrect, and errors exist in the labeling process, which imposes great noise to the model training. These errors primarily consist of wrong labeling information, excessive labeling and missing labeling, resulting in only 71.6% of the original mAP of the model. Moreover, the mAP of the proposed model was enhanced evidently after these erroneous data were manually corrected.

To achieve more effective experimental results, other methods are adopted as well. Color acts as a vital feature in goods recognition. The colors of the three channels are uniformly distributed in the image, the colors are normalized to a uniform scale, and the mean value of each channel is set to (127,127,127); subsequently, the model is trained again. However, it is identified that the effect on the results of the model is not significant, and the accuracy remains about 95%, that means color normalization does not help improve the detection precision, this method is not adopted in this study.

#### 3) RECOGNITION OF SPECIAL ITEM TYPES

As can be seen from Figure 7, the item recognition rate of class 16 is relatively low, the recognition rate is only 84.2%. This is an extreme case in a practical scenario is illustrated in Figure. 9, the top row and the bottom row of items in the red box have almost the same characteristics of bottle caps, but they are actually two different commodities with different prices. For the insufficient bottle characteristics, these two types of goods are difficult to distinguish only from the bottle cap feature. The proposed model can detect the position of these goods, but can not be very accurate classification. The similar placement as shown in Figure. 9 is avoided in the present practical application.

**FIGURE 9.** A practical scenario of items special placement.

## V. CONCLUSION

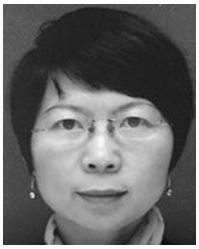
In the present study, a smart unstaffed retail shop scheme is proposed based on AI and the IoT, as an attempt to verify the feasibility of the unstaffed retail shopping style implementation and solve the unobvious object features attributed to the mutual occlusion of items in the vending machine. A SSD (300 × 300) model with sub-prediction structure is designed to analyze multiple target features of commodities, the mAP of the proposed SSD (300 × 300) model reaches 96.1%, and the experimental results reveals that the proposed

SSD (300×300) model outperforms than the Fast R-CNN and Faster R-CNN in detection. The proposed SSD (300 × 300) model also has significantly higher detection precision than the original SSD (300 × 300) model while the detection time increased by only 1ms. The proposed system can satisfy the requirements of real applications. The customer flow and the transactions of the shops in new retail mode based on the designed system are increased significantly. In the subsequent work, we will focus on further improving the algorithm recognition rate and efficiency by adjusting model structure and model design.

## REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.
[2] X. Lin, J. Li, J. Wu, H. Liang, and W. Yang, "Making knowledge tradable in edge-AI enabled IoT: A consortium blockchain-based efficient and incentive approach," *IEEE Trans. Ind. Informat.*, vol. 15, no. 12, pp. 6367–6378, Dec. 2019.
[3] J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, "Big data analysis-based secure cluster management for optimized control plane in software-defined networks," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 1, pp. 27–38, Mar. 2018.
[4] A. Khanna and R. Tomar, "IoT based interactive shopping ecosystem," in *Proc. 2nd Int. Conf. Next Gener. Comput. Technol. (NGCT)*, Oct. 2016, pp. 40–45.
[5] J. Rezazadeh, K. Sandrasegaran, and X. Kong, "A location-based smart shopping system with IoT technology," in *Proc. IEEE 4th World Forum Internet Things (WF-IoT)*, Feb. 2018, pp. 748–753.
[6] Q. Yi and P. Li, "Design and implementation on supermarket shopping guide system based on RFID and Internet of Things," *J. Comput. Res. Develop.*, vol. 47, no. 2, pp. 350–354, 2010.
[7] *Inside Amazon Go: The Store of the Future*. Accessed: Jan. 26, 2018. [Online]. Available: https://money.cnn.com/2018/01/26/technology/amazon-go-store/index.html
[8] I. Tam. (Jul. 2018). Alibaba unveils staff-less tao cafe and smart speaker to revolutionise offline retail. Marketing Interactive. [Online]. Available: https://www.marketing-interactive.com/alibaba-unveils-staff-less-taocafe-and-smart-speaker-to-revolutionise-offline-retail/
[9] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," in *Proc. CVPR*, 2019, *arXiv:1901.07249*. [Online]. Available: https://arxiv.org/abs/1901.07249
[10] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*. [Online]. Available: https://arxiv.org/abs/1905.05055
[11] S. Agarwal, J. Ogier Du Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," 2018, *arXiv:1809.03193*. [Online]. Available: http://arxiv.org/abs/1809.03193
[12] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, 2015, pp. 1440–1448.
[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[16] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[17] J. Wu, M. Dong, K. Ota, J. Li, W. Yang, and M. Wang, "Fog-computing-enabled cognitive network function virtualization for an information-centric future Internet," *IEEE Commun. Mag.*, vol. 57, no. 7, pp. 48–54, Jul. 2019.
[18] J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, "FCSS: Fog-computing-based content-aware filtering for security services in information-centric social networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 7, no. 4, pp. 553–564, Oct. 2019.
[19] S. Qiao, W. Shen, W. Qiu, C. Liu, A. Yuille, "Scalenet: Guiding object proposal generation in supermarkets and beyond," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1791–1800.
[20] B. Hu, N. Zhou, Q. Zhou, X. Wang, and W. Liu, "DiffNet: A learning to compare deep network for product recognition," *IEEE Access*, vol. 8, pp. 19336–19344, 2020.
[21] L. Liu, B. Zhou, Z. Zou, S.-C. Yeh, and L. Zheng, "A smart unstaffed retail shop based on artificial intelligence and IoT," in *Proc. IEEE 23rd Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Barcelona, Spain, Sep. 2018, pp. 1–4, doi: 10.1109/CAMAD.2018.8514988.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[23] Z. Cai, Q. Cai, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 9908, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-46493-0_22.
[24] K. Wongsuphasawat, D. Smilkov, J. Wexler, J. Wilson, D. Mané, D. Fritz, D. Krishnan, F. B. Viégas, and M. Wattenberg, "Visualizing dataflow graphs of deep learning models in tensorflow," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 1–12, Jan. 2018.
[25] H.-Y. Zhou, A. Oliver, J. Wu, and Y. Zheng, "When semi-supervised learning meets transfer learning: Training strategies, models and datasets," 2018, *arXiv:1812.05313*. [Online]. Available: http://arxiv.org/abs/1812.05313
[26] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

**JIANQIANG XU** (Member, IEEE) received the M.B.A. degree from the East China University of Science and Technology, in 2009, where he is currently pursuing the Ph.D. degree. His research interests include the Internet of Things, artificial intelligence, and cloud computing.



**ZHUJIAO HU** received the M.Sc. degree in engineering from Wuhan University, in 2005. She is currently pursuing the Ph.D. degree with Fudan University. Her research interests include the Internet of Things, artificial intelligence, and big data.

**ZHUO ZOU** (Member, IEEE) received the Ph.D. degree in electronic and computer systems from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2012.

He was with the iPack VINN Excellence Center, KTH, as the Assistant Director, where he coordinated the research project on ultra-low-power embedded electronics for wireless sensing. He currently holds the Outstanding Young Professorship in the School of Information Science and Technology, Fudan University, Shanghai, China. His current research interests include ultra-low power circuits and systems, energy efficient SoC design, and smart systems for the Internet of Things.

**JUNZHONG ZOU** received the B.S. degree in electrical engineering from Chongqing University, Sichuan, China, in 1982, and the M.S. and Ph.D. degrees in advanced systems control engineering from Saga University, Japan, in 1995 and 1998, respectively.

He joined the East China University of Science and Technology, Shanghai, China, in 2001. He is currently a Professor and Thesis Adviser of M.S. and Ph.D. graduate students in ECUST. He is also the Director of the Texas Instrument Digital Signal Processing Associated Lab, ECUST. His research interests are dynamical control and automation, biomedical signal processing, robotics, and mechatronic servo systems.

**XIAOMING HU** (Senior Member, IEEE) received the B.S. degree from the University of Science and Technology of China, in 1983, and the Ph.D. degree from Arizona State University, in 1989, under the guidance of Professor Christopher I. Byrnes. He served as a Research Assistant at the Institute of Automation, Chinese Academy of Sciences, from 1983 to 1984. From 1989 to 1990, he was a Gustafsson Postdoctoral Fellow with the KTH Royal Institute of Technology, where he has been a Full Professor of optimization and systems theory, since October 2003. His main research interests are nonlinear systems, multi-agent systems, active sensing and nonlinear observers, and optimal control.

**LIZHENG LIU** (Member, IEEE) received the B.S. degree in electronic engineering and the M.S. degree in automation from Xiangtan University, Xiangtan, China, in 2002 and 2006, respectively, and the Ph.D. degree in microelectronics and solid state electronics from Fudan University, Shanghai, China, in 2019. He holds a postdoctoral position at Fudan University. He is a Guest Researcher with KTH, Sweden. His researches mainly focus on brain-like computing, artificial intelligence (AI), and smart hardware.

**LIRONG ZHENG** (Senior Member, IEEE) received the Ph.D. degree in electronic system design from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2001. Afterward, he worked at KTH as a Research Fellow, an Associate Professor, and a Full Professor. He is the Founding Director of the iPack VINN Excellence Center, Sweden, and the Chair Professor in media electronics with KTH, since 2006. He is also a Guest Professor since 2008, and a Distinguished Professor since 2010, with Fudan University, Shanghai, China. He currently holds the Directorship of the Shanghai Institute of Intelligent Electronics and Systems, Fudan University. His research experiences and interests include electronic circuits, wireless sensors and systems for ambient intelligence, and the Internet of Things. He has authored more than 400 publications and serves as steering board member of International Conference on Internet-of-Things.