Locating in Supermalls with SSD

Yuheng Ding

Nanjing Foreign Language School Nanjing, China

18801589870@163.com

Keywords: Indoor positioning method, convolutional neural network, Single Shot Multi-Box Detector (SSD)

Abstract: The need for a fast and accurate indoor positioning solution is increasing rapidly since people are spending more time indoors, mostly for shopping and entertaining. Being aware of their own positions is not only vital to consumers' safety, but also necessary to improve their shopping experience. In this paper, an alternative way for indoor positioning is proposed, which is based on convolutional neural network, particularly an algorithm called Single Shot Multi-Box Detector (SSD). Different from other indoor positioning technologies, this image-based technique does not depend on special detecting devices or communication networks. Testing shows that positioning with SSD in indoor spaces has great advantages (both convenience and accuracy) over other methods, indicating its potential to be applied in real-life situations.

1. Introduction

With the rise of various real-time locating systems, people's need to get hold of their own positions increase rapidly. For outdoor activities, GPS can effectively give people their precise location. Being all-dimensional, all-weather, and all-time, GPS makes positioning and navigation possible. However, GPS is not proper for all circumstances. At present, the deviation of GPS for civilian uses is around 10 meters, indicating its inappropriateness for smaller cases, such as supermalls. Nowadays, people spend nearly 70% of their time indoors. Statistics show that a large amount of consumption takes place in places such as supermalls. To improve consumers' shopping experience, a location instruction with the variation less than 2 meters is necessary. When people enter an unfamiliar supermall, they are likely to get lost because of the huge size and the complex layout. The traditional guideposts can hardly provide accurate position for everyone. Indoor positioning system can assist the customers to find their locations and find where they want to go. In crowded indoor spaces, companions may get lost from each other. Parents will be anxious if they cannot find their kids. With indoor positioning systems, parents can know where their children are at any time. Additionally, when emergency happens indoors and people are trapped in supermalls or buildings, indoor positioning systems can help rescuers to locate those trapped, improving the rescue efficiency and striving for the precious time.

Indoor positioning based on computer vision have gained great advance recently. Attempts to use Augmented Reality and Virtual Reality have been made to help positioning. But they both need head-worn display equipment which is not convenient for customers. Besides, the amount of calculations for reconstruction is too large for a portable device. There are also ways using visual markings such as QR codes. But QR code based methods require labeling all over the places, causing esthetic problems. The traditional way to retrieve images is the Text Based Image Retrieval. It relies on textual markings on images and retrieval of texts. As a mature retrieving technique, it mainly utilizes human's understanding of natural languages. This method needs textual description by artificial or semi-supervised participation, which includes the theme, location, figures in the images. Generally, it transfers image information to textual information and matches the search content. However, this original artificial way costs much time. The efficiency of this method cannot be guaranteed.

In order to make more precise positioning, we propose an alternative way of using phones to recognize brands of the shops which are connected with position information. It is based on the SSD algorithm built on the deep learning framework Caffe. Through Python programming, it can achieve the goal of identifying the brands in the input image and then output the location information. Table 1 shows a comparison of various indoor positioning methods.

Single Shot Multi-Box Detector [1] is an algorithm for detecting objects in images using a single deep neural network. SSD uses a small convolutional filter to predict object categories and offsets in bounding box locations. Different from other object detection systems which resample pixels or features for bounding box hypotheses, SSD produces predictions of different scales from feature maps of different scales, and explicitly separate predictions by aspect ratio. These improvements lead to higher efficiency while remain the accuracy. There are many other existing methods for image recognition. YOLO [2], for example, frames object detection as a regression problem to spatially separated bounding boxes and associated class probabilities, which can be optimized end-to-end directly on detection performance. However, because it makes

Table.1. Comparison of different positioning methods

Positioning Methods	Comparison Features				
	Accuracy	Scope	Device	Cost	Ease of Use
RFID	High	Large	System server & card reader	Medium	Additional devices needed
WiFi Positioning	Low	Medium	Cell phone	Low	Easy
Bluetooth	High	Medium	Bluetooth beacon & cell phone	High	Easy
Ultrasonic Wave	High	Small	Ultrasonic wave sensors	Medium	Additional devices needed
SSD Positioning	High	No Limits	Cell phone	Low	Easy

Some localization errors, so it fails to accomplish my goal. On the contrary, SSD adds several feature layers to the end of the base network, increasing the accuracy while maintaining the speed. Another pervasive method for object detection is R-CNN [3]. It extracts bottom-up region proposals and computes features with a large convolutional neural network. Then it classifies each region using class-specific linear SVMs. In general, R-CNN achieves a great accuracy. But it is very computationally intensive, which decreases the recognition speed.

As our aim is to recognize the picture instantly, we intend to choose a method with high speed and fair accuracy. The comparison between different recognition algorithms shows that SSD is less computationally intensive, contributing to its fast speed. Its fair accuracy also indicates that high speed does not come at the cost of significantly decreased detection accuracy.

2. Data Processing

2.1 Data Acquisition

The data used in this model come from videos of 20 brands of shops' in a supermall. The brands of the shops are chosen randomly to avoid bias. In other words, the shops are located in all levels of the supermall. We shoot the videos while walking slowly in front of each shop in order to get a clear image of the brand. The lengths of the videos are all precisely set to 10 seconds. During the 10 seconds, we walked with an average speed of 0.6 meters per second, so the videos can cover the whole views of the brands.

2.2 Data Size

Every second the camera films 29.84 frames, so it is estimated that every video generates 300 pictures. The total amount of the original data is around 6,000. Pictures from the same video are considered to be from the same class, thus 6,000 pictures are divided into 20 classes in total. 60% of the pictures are used to train the model while the rest are used for testing.

2.3 Annotated Attributes

In order to train the model and evaluate the classification and detection, we add annotations to help the process [4]. The annotation attached to each picture includes two parts: (i) class (one of the 20 brands represented by single letters such as a, b, c); (ii) bounding box (a rectangular box surrounding the extent of the brand visible in the image). Each bounding box is drawn by hand to locate the brand in the image. Fig.1 shows two samples of data with annotations.

2.4 Data Augmentation

Because of insufficient amount of training data, the increment of the data size is necessary for a comprehensive model. We generally adopt the method used in the SSD. In addition to the original extracted pictures with full size, these pictures are also randomly transferred to different scales. The Patch can either has the minimum Jaccard number overlap with the objects of 0.1, 0.3, 0.5, 0.7, 0.9 or be randomly sampled. In general, the patch after transformation will be smaller or have the same size with the original picture, and the aspect radio is between 0.5 and 2.

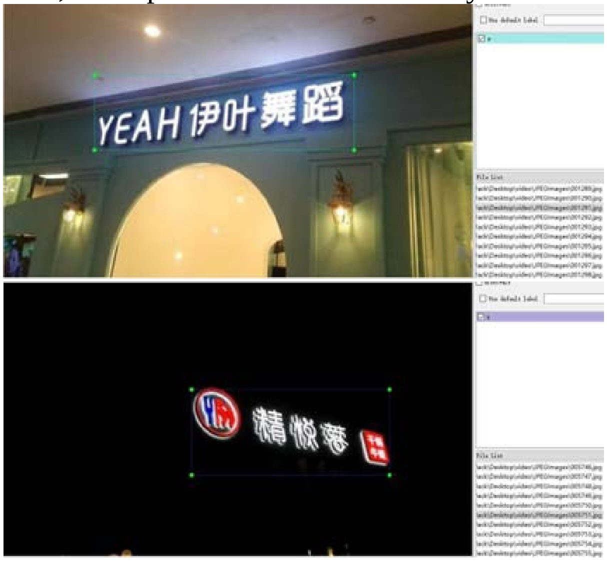


Figure 1. Two samples of data with annotations.

3. SSD algorithm

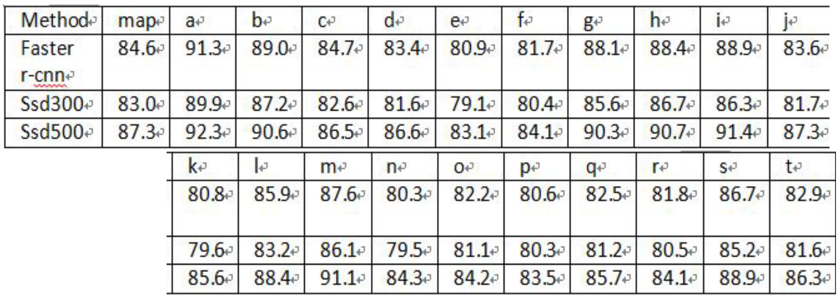
Unlike other detection approach, SSD [1] only needs an input image and ground truth boxes for each object during training. It evaluates a small set of default boxes with different aspect ratios at each location in several feature maps with different scales. When training, SSD matches default boxes to the ground truth boxes.

SSD uses VGG-16 as its base. VGG-16 [5] has 16 layers, including 13 convolutional layers and 3 fully connected layers. Compared to other state-of-the-art networks, VGG-16 can effectively reduce the error rate. It also has great scalability, which means it can be generalized to various forms. It generally makes use of convolutional kernels of size 3*3 and pooling kernels of size 2*2. Its performance can be improved by deepening the network and adding more layers.

SSD adds convolutional feature layers to the end of the truncated base network. Such layers have their sizes decreasing, allowing predictions of detections at multiple scales. Each added feature layer can produce a constant set of detection predictions using a set of convolutional filters. The training material of the SSD is default boxes in correspondent with a ground truth detection. Default boxes selected according to each ground truth box vary over location, aspect ratio, and scale.

4. Testing Results

We compared SSD with Faster R-CNN in 20 classes, obtaining the results shown in Fig. 2. SSD with 512 pixel outperformed Faster R-CNN in all classes. On the contrary, SSD with 300 pixels has lower mean average precisions due to its lower pixel. This means that the resolution of the picture has a major influence on the precision of the result that cannot be neglected. Fig.3 shows the testing interface.



Method	map	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
Faster r-cnn	84.6	91.3	89.0	84.7	83.4	80.9	81.7	88.1	88.4	88.9	83.6	80.8	85.9	87.6	80.3	82.2	80.6	82.5	81.8	86.7	82.9
Ssd300	83.0	89.9	87.2	82.6	81.6	79.1	80.4	85.6	86.7	86.3	81.7	79.6	83.2	86.1	79.5	81.1	80.3	81.2	80.5	85.2	81.6
Ssd500	87.3	92.3	90.6	86.5	86.6	83.1	84.1	90.3	90.7	91.4	87.3	85.6	88.4	91.1	84.3	84.2	83.5	85.7	84.1	88.9	86.3

Figure 2. Comparing SSD with Faster R-CNN.

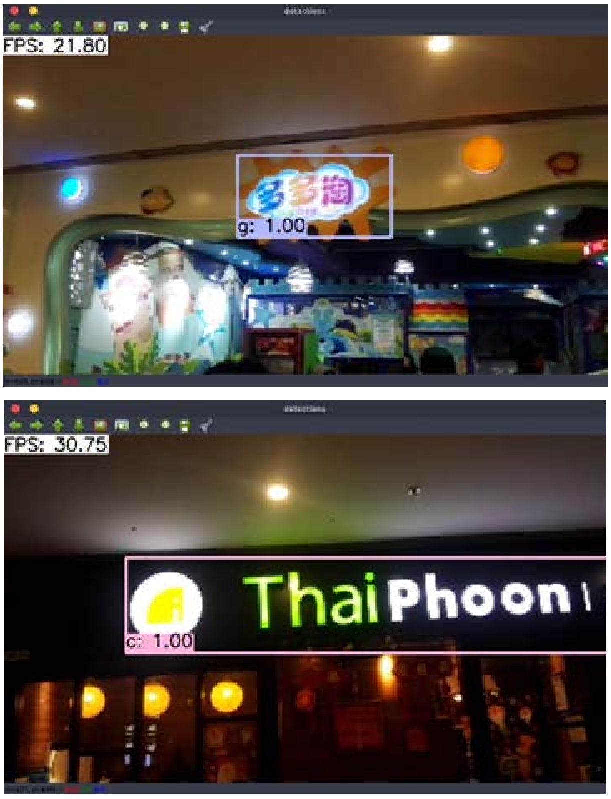


Figure 3. Testing interface.

5. Application

As shown in Fig. 4, for a specific application scene, the user first uses a cell phone or a camera to take the picture of a brand. Then the picture will be sent to the server automatically. Inside the server, GPUs will compute the picture with SSD algorithm, during which the brand will be located and classified. As the class which the picture belongs to is connected directly with the location information, the location will be determined. But in some special cases (for example, the camera is carried by a child or a pet), the information will be sent to another receiver, the parent or the owner.

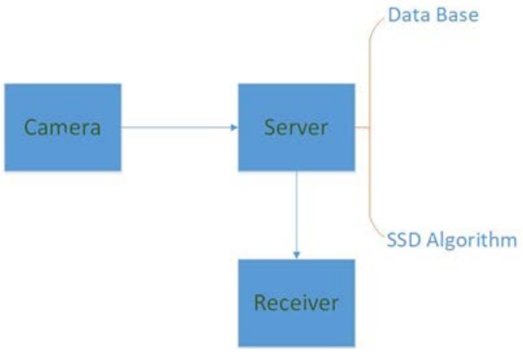


Figure 4. Application diagram.

6. Conclusion

In this paper, we propose an alternative way of positioning in indoor space with image recognition technology. Among various indoor positioning methods using ultrasonic wave, Wi-Fi and Bluetooth, our method outperforms other indoor positioning technologies by accuracy, cost, and ease of use. Furthermore, we adopt Single Shot Multi-Box Detector, which stands out for its high speed and acceptable accuracy: a cost-effective method. The prospect of the practical use of our method is enormous. People can use it to locate themselves or to find their children and pets.

There are also a few aspects for future improvement. Our method needs a server with numerous GPUs. The prices of the GPUs are expensive. Also, if too many people call for their location at the same time, the speed will not be so fast and the server may not be able to deal with their request immediately. Besides, if there are a great number of shops, which means there are too many classes, the accuracy of the recognition will decrease and it will cost more time. In order to employ this method to a large supermall, improvement must be made to ensure that the accuracy still remain acceptable.

References

[1] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector [C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[2] Redmon J, Farhadi A. YOLOv3: An incremental improvement [J]. arXiv preprint arXiv: 1804.02767, 2018.

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[4] Everingham M, Eslami S M A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective [J]. International journal of computer vision, 2015, 111 (1): 98-136.

[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.