Mathematical Biosciences and Engineering

MBE, 17(1): 266–285.
DOI: 10.3934/mbe.2020015
Received: 30 June 2019
Accepted: 25 September 2019
Published: 08 October 2019

http://www.aimspress.com/journal/MBE

Research article

An intelligent indoor positioning system based on pedestrian directional signage object detection: a case study of Taipei Main Station

Chun-Chao Yeh*, Ke-Jia Jhang and Chin-Chun Chang

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung 20224, Taiwan

* **Correspondence:** Email: ccyeh@mail.ntou.edu.tw; Tel: +886224622192 ext. 6604; Fax: +886224623249.

Abstract: Indoor positioning technologies have gained great interest from both industry and academia. Variety of services and applications can be built based on the availability and accessibility of indoor positioning information, for example indoor navigation and various location-based services. Different approaches have been proposed to provide indoor positioning information to users, in which an underlying system infrastructure is usually assumed to be well deployed in advance to provide the position information to users. Among many others, one common strategy is to deploy a bunch of active sensor nodes, such as WiFi APs and Bluetooth transceivers, to the indoor environment to serve as reference landmarks. The user's current location can thus be obtained directly or indirectly according to the active sensor signals collected by the user. Different from conventional infrastructure-based approaches, which put additional sensor devices to the environment, we utilize available objects in the environment as location landmarks. Leveraging wildly available smartphone devices as customer premises equipment to the user and the cutting-edge deep-learning technology, we investigate the feasibility of an infrastructure-free intelligent indoor positioning system based on visual information only. The proposed scheme has been verified by a real case study, which is to provide indoor positioning information to users in Taipei Main Station, one of the busiest transportation stations in the world. We use available pedestrian directional signage as location landmarks, which include all of the 52 pedestrian directional signs in the testing area. The Google Objection Detection framework is applied for detection and recognition of the pedestrian directional sign. According to the experimental results, we have shown that the proposed scheme can achieve as high as 98% accuracy to successfully

identify the 52 pedestrian directional signs for the three test data sets which include 6,341 test images totally. Detailed discussions of the system design and the experiments are also presented in the paper.

Keywords: indoor positioning; signage detection; deep-learning; R-CNN; smart environment

1. Introduction

We have experienced convenience and benefits of versatile outdoor positioning services such as location positioning, navigation, tracking, and various location-based services. Most of these outdoor positioning services are based on a global positioning infrastructure provided by ubiquitous GNSS (Global Navigation Satellite System) services such as GPS(US), GLONASS(Russia), BDS(China), and Galileo(EU). Since GNSS services are based on receiving signals from satellites, they are invalid at the place where satellite signals cannot reach, for example inside a building. Consequently, indoor positioning has been a hot research topic. Indoor positioning technologies have gained great interest from both industry and academia. Variety of services and applications can be built based on the availability and accessibility of indoor positioning information, for example indoor navigation and various location-based services. Different approaches have been proposed to provide indoor positioning information to users, in which usually an underlying system infrastructure is assumed to be well deployed in advance to provide the position information to users. Among many others, one common strategy is to deploy a bunch of active sensor nodes, such as WiFi APs and Bluetooth transceivers, to the indoor environment to form a sensor network, in which each of the sensor node serves as a reference landmark. Via referring to sensor locations as landmarks, the current location of a user can be obtained directly or indirectly according to the sensor signals from different sensor nodes, collected by the user at the current location. Nonetheless, there are two main concerns with such infrastructure-based indoor positioning systems: deployment cost and accessibility.

Most of indoor positioning systems rely on interactions between active sensor nodes (e.g. WiFi Access Point, Zigbee node, and RFID) and their corresponding receiver/tag devices carried by users. For example, most WiFi-based positioning systems use the fingerprints of RSSI (receiving signal strength indication) and the SSID (Specific Service Set Identifier) information detected by the user WiFi device. A higher RSSI value received by a user to a WiFi access point indicates the user is closer to the access point. Combining the RSSI signals receiving from multiple WiFi access points nearby, the current location of a user can be approximately derived based on the predicted distances between the user and the WiFi access points nearby. Similar mechanisms can be applied to the cases of Zigbee and RFID. However, availability of user devices should be taken into account in practice. For example, among the four sensor devices aforementioned: WiFi, Bluetooth, Zigbee, and RFID, the cases for WiFi and Bluetooth are more feasible as they are widely available in smartphones. In contrast, the cases for Zigbee and RFID are more restricted since most of people do not have Zigbee/RFID receivers/tags in hands at any time.

Different form conventional infrastructure-based approach, which put additional sensor devices to the environment, we try to utilize available objects in the environment as location landmarks. Leveraging wildly available smartphone devices as customer premises equipment to the users and the cutting-edge deep-learning technology, we investigate the feasibility of an infrastructure-free intelligent indoor positioning system based on visual information only. The proposed scheme does

not require to deploy any active sensor nodes precisely, which is important in practices. It is not only due to the reason of avoiding high deployment cost but also due to time and other constraints for realization. The proposed scheme is verified by a real case study, which is to provide indoor positioning information to users in Taipei Main Station, one of the busiest transportation stations in the world. In this case study, we use available pedestrian directional signage as location landmarks. Via a pre-download smartphone App, a user can get his/her current location by uploading an image of a nearby pedestrian directional sign to the positioning system. Then, the system identifies the pedestrian directional sign in the uploaded image and indicates the current location of the user. All of the 52 pedestrian directional signs in the testing area are adopted for this study. The content of each pedestrian directional sign consists of three parts in general: the street/building/facility name both in Chinese and English, and the icon for the direction sign and the facility. The Google Objection Detection framework is applied for detection and recognition of pedestrian directional signage. According to the experimental results, we have shown that the system can achieve as high as 98.3% (0.982930) accuracy in average to successfully identify the correct one among the 52 pedestrian directional signs for the three test data sets which include 6,341 testing images acquired from different cameras and users. Detailed discussions of the system design and the experiments are presented in the paper.

The remainder of the paper is organized as follows. In Section 2, we review some of previous research results related to our works. In Section 3, we present the rationale and the architecture framework of the proposed indoor positioning system. In Section 4, we depict the design of a high accurate signage image object detection subsystem. Results and discussions of the performance evaluation on the signage object detection are presented in Section 5. Finally, a brief conclusion is given in Section 6.

2. Related works

2.1. Indoor positioning technologies

One of popular indoor positioning technologies is based on WLAN signal fingerprints, in which user's current location is estimated by a set of RSS values received from neighboring WLAN APs (access points) referred as RSS fingerprints [1–3]. It is a purely data-driven method based on statistics of RSS fingerprints at each of reference location points. The approach relies on a preinstalled WLAN infrastructure, and a deliberate survey of RSS fingerprints at each of reference locations. Since the approach is based on the RSS radio signal, it is sensitive to signal quality due to the poor stability of WIFI signals. In a complicated indoor environment such as public hot spots, visitors come and go and the population of visits changes dynamically. Under such a circumstance, it is hard to construct a sophisticated RSS fingerprint database to precisely estimate user's current location based on the RSS values received by the corresponding end device carried by the user. Ifthekhar etc. [4] proposed a novel IPS scheme, which is based on optical camera communication (OCC) infrastructures. The proposed scheme relies on an OCC environment where a set of well-setup LEDs mounted on the ceiling provide the light source required by the OCC scheme. The authors assume the positions of these LEDs are known, and the coordinate information of each LED is delivered to user cameras via OCC mechanisms. The world coordinate of the camera is estimated via the mapping between the coordinates of the LEDs shown in an image taken by the camera. The

author proposed to use a neural network to solve the complicated relationship between the 3D coordinate information (world coordinates of the LEDs mounted on the ceiling) and the 2D coordinate information of these LEDs in the image. Since LED-based OCC is not so popular in real-world deployment, feasibility of the proposed OCC-based scheme is questionable. Another problem is that at least three LEDs must be captured simultaneously by the smartphone camera and the smartphone should be face-up to the ceiling while taking the picture. This implies that the LEDs need to be densely deployed. Meanwhile, it is hard to use in a crowded environment as the use case which we focus on.

In [5], the authors proposed a real-time indoor positioning system based on UWB signals. The proposed scheme relies on TDOA (time difference of arrival) information of multiple UWB (ultra-wideband) stations. Traditionally, UWB TODA approach is sensitive to multiple-path signal degradation and NLOS (non-line of sight) situations. To reduce the effects, multiple UWB channels are used. All base stations use four transceivers instead of one to increase the system robustness. Like all the schemes based on TDOA, the accuracy of the proposed scheme relies on high precision of time synchronization between all the transceivers. Also, the approach is more sensitive to a crowded environment, in which the multiple path and NLOS effects are harsh. Gan et al. [6] investigated an indoor positioning system, which combines pseudolite (a satellite-like GPS ground signal generator) and PDR (Pedestrian Dead Reckoning) technologies. The pseudolite devices play a role as indoor GPS transceivers. However, one of the major downside of using pseudolite devices is installation cost. Moreover, the pseudolite signals are prone to be blocked by complicated indoor objects, such as walls and rooms. Consequently, the authors suggested to combine the usage of the PDR scheme. The PDR devices monitor the user's movement on a 2D plane. By measuring the user's walking steps and directions, the PDR device estimates the user's current position on the 2D plane, if the user's initial position and azimuth are known. Compared with the pseudolite device, the PDR device costs less. However, PDR approaches have error accumulation problems. It needs to be calibrated frequently by other precise IPS schemes to reduce the accumulated estimation errors.

2.2. Convolutional neural networks (CNNs)

Current state-of-the-art object detection systems utilize convolutional neural networks (CNNs), which are multi-layer neural networks with shared weights and local receptive fields. These CNNs usually have deep network architectures, comprise of millions of parameters, and are able to represent images in various semantic levels. In general, the state-of-the-art convolutional object detection system belongs to one of the two approaches [7]: one-stage detectors and two-stage detectors. The two-stage detector, such as R-CNN [8], Faster R-CNN [9], R-FCN [10], has two consecutive stages. In the first stage, candidate object locations are generated by algorithms of generating categorical-independent region proposals, such as selective search [11], or by region proposal networks [9]. In the second stage, CNNs are employed to classify each region proposal as one of the object of interest or as the background. The one-stage detector, such as YOLO [12], SSD [13], RetinaNet [14], has a single network architecture for localizing and identifying multiple objects in an image. The one-stage detector is often faster than the two-stage detector, whereas the two-stage detector is often more accurate than the one-stage detector. In [7], it turns out that there is a trade-off between the accuracy, speed, and memory usage in state-of-the-art convolutional object detection systems. For a given application on a specific platform, the network architecture for the

convolutional object detector can be chosen by balancing the accuracy, speed, and memory space of the detector. In addition, fine tuning the pre-trained model of a convolutional object detector with a few labeled images of target objects is an economic way to apply the object detector to a specific application. There are such open source frameworks of object detectors, such as YOLO [15] and the Tensorflow Object Detection API [16]. In this paper, the Tensorflow Object Detection API is adopted.

(a) Taipei Main Station Area 3D map (Picture source: http://weijidai.pixnet.net/album/set/15840742)

(b) Taipei Main Station information map (B1 Area)

Figure 1. Taipei Main Station(TMS). (a) a 3D map of TMS; (b) a visitor information map for B1 Area.

3. Demands and proposed system framework

3.1. Demands

Our intention for this research study is to develop a positioning system to help internet users to acquire position information of the location where the users currently are. In particular, in this research study, we develop an effective scheme to provide an indoor positioning system for visitors/passengers in a large train/subway station such as Taipei Main Station (TMS) [17] in Taipei, Taiwan. TMS is the biggest station in Taiwan. It is a building complex interconnecting up to five transportation terminal stations (Taiwan Railways Taipei Station, Taiwan High Speed Rail Taipei Station, Taipei Metro, Taoyuan Airport MRT, and the Taipei Bus Station) through underground passageways. According to a survey in 2018 by Taipei City Government, the annual passenger served in 2018 by Taipei Main Station is approaching 190 million (189, 952, 804) [18], including both of entries (95, 204, 548) and exits (94, 748, 256). The number are contributed by passengers for Taiwan Railways Taipei Station (44, 077, 231), Taiwan High Speed Rail Taipei Station (30, 402, 961), and Taipei Metro TMS station (115, 472, 612). Regarding the passenger population, TMS serves even more passengers than London Waterloo Station (around 100 million passengers in the year of 2017–2018 [19]), one of the largest stations in Europe.

The building complex of TMS connects not only stations but also Taipei City Mall. Figure 1a

shows a 3D-view graph of TMS. There are more than fifty entrances/exits connecting the building complex and surrounding areas via underground passageways. For most of visitors to TMS, the underground of TMS seems to be a maze. It is prone to get confused for which ways/direction they should take to reach their destination. While it might be helpful to visitors if the organization/institute can offer digital map and/or floor plan information via internet, it would be much better if the system can provide visitor position information to let the visitors know where they are right now. Moreover, with enabling the position information, many location-based services can be offered.

Accordingly, in this research study, an effective scheme is developed to provide an indoor positioning system for visitors/passengers in a large station such as Taipei Main Station. To provide a cost-effective positioning system, we consider following system demands.

(1) High accessibility: Most of popular indoor positioning architectures are based on sensor signals collected from a set of well-deployed sensor devices. Depending on the system architecture, usually different types of user premise equipment are required to be available to users to receive the sensor signals. While it is possible to have the special user premise equipment from the service providers, it would be inconvenient to users due to the extra administrative procedure to borrow the equipment. Instead, ideally, such user premise equipment should be something available to users at hand, such as smartphones. Consequently, anyone can easily access the positioning service.

(2) High reachability: According to the acquisition mechanism of the position sensing information and the positioning system deployment scheme, there are different degrees of limitations or assumptions to get the positioning information. For example, for a positioning system relies on a passive sensor tag such as QR-codes or RFID tags, usually users should get very close to the sensor tag to acquire the positioning information. For a crowded indoor environment such as a busy train station, it could not be easy to find a clear path to reach the sensor tag close enough. Another practical concern is that usually it is not so easy for users to find the places the sensor tags located.

(3) Low deployment cost. While it is not an essential technical issue, deployment cost should be taken into account in practice. At least two costs are needed to be considered: capital expenditures (Capex) and operational expenditure (Opex). The Capex includes all of the software and hardware cost of the sensing infrastructure enabling the positioning service such as the location information sensors/transmitters in the front-end and the data processing servers in the back-end. The Opex includes the energies and the human resources to keep the service functions available all the time.

(4) Low deployment limits. Another practical concern is the environment constraints on the deployment of the positioning system. For example, if the deployment involves to install massive active sensors (such as the WiFi AP, RFID transceiver, and Bluetooth iBeacon transceiver, to name a few) to cover all the building complex (both internal and external), both sensor device installation and related wiring (for data and power transmission) issues might be a problem due to some concerns such as aesthetic value, security, and safety.

(5) Robust signal quality. Many indoor positioning systems rely on accurate radio signals received from reference sites. While these systems might work perfectly in an ideal environment, where no much radio interference is presented, they become less accurate as the place is getting crowded. Due to the nature of radio signal propagation, in a crowded place, radio signals would

be blocked by the crowd, which could cause severe multipath problems to the radio signal and thus degrade the signal quality. Accordingly, how to provide a reliable indoor positioning system in a crowded place becomes a challenging problem.

(6) Easy to use. Last but not least, for users, it should be easy to use. No long learning curve is expected.

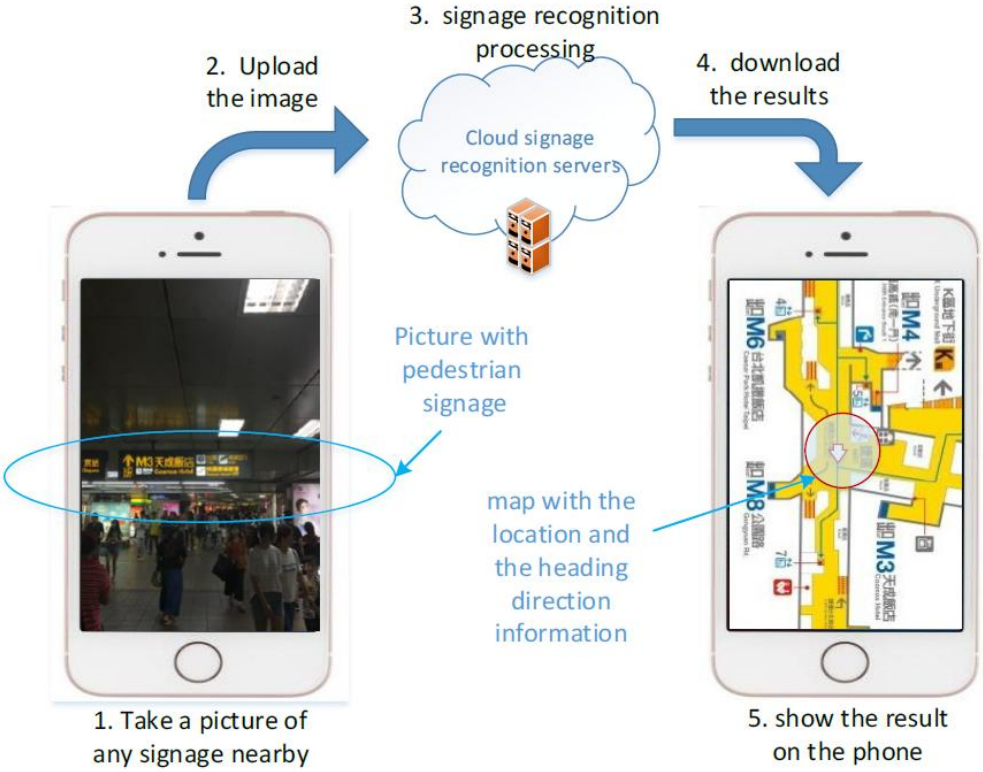

Figure 2. Proposed system framework for indoor positioning services based on pedestrian signage recognition.

3.2. Proposed system framework

In this research study, we propose a new design strategy to provide indoor positioning services based on the demand for a large train station such as Taipei Main Station, in which we propose to use signage recognition as the basis for the indoor positioning service. The general idea of the proposed scheme is as shown in Figure 2. We assume that users have a smartphone in hands and they are able to connect to internet via mobile telecommunication networks (e.g. 4G LTE mobile networks) or WiFi wireless LANs. To access the positioning service, users first need to install the provided mobile application software (mobile App) for the positioning service. Taking the TMS scenario as the use case we investigate in this research study, when a visitor in TMS would like to check his/her current location (and the heading direction) inside TMS, he/she just needs to use the mobile App to take a picture of any pedestrian directional signage near to him/her (Step 1 in Figure 2). The mobile App then will upload the picture to the backend server for signage object detection and recognition (Steps 2 and 3). The pedestrian directional sign plays a role as a location landmark. Via signage object detection and recognition, the system identifies which pedestrian directional sign is shown in the

image. The location of the identified pedestrian directional sign is marked on the map and sent back to the user's smartphone (Steps 4 and 5) as shown in Figure 2.

Table 1. Comparison: indoor positioning services in crowd indoor field area.

method	WiFi	BLE iBeacon	Zig-bee	RFID-tag	QR-code	proposed scheme*
Type	active	active	active	passive	passive	passive
High accessibility	++++	++++	+	++(++)	++++	++++
High reachability	++++	+++	++	+	+	++++
Low deployment cost	+	+	+	++	++	++++
Robust signal quality	+	+	+	++	++	++++
Easy to use	++++	++++	+	++++	++++	++++
precision	++	+++	++	++++	++++	++

*Proposed scheme: signage recognition based.

Compared with those approaches based on radio signals (such as WiFi, BLE iBeacon, and Zig-bee), the proposed scheme is easy to use and provides high availability and accessibility. Moreover, the deployment cost is nearly free, and it is more robust to a large number of visitors/passengers such as for the case as in a busy train station. In Table 1, we make a comparison between the proposed scheme and some other popular approaches for indoor positioning services. The comparison is based on the design considerations aforementioned. Regarding accessibility, our proposed scheme uses smartphones as user premise equipment, which are popular especially for young generations in most of countries. Similarly, for the QR-code approach, most of smartphones can install a QR-code reader App and it would be no problem to users to use their smartphones as QR-code readers. Meanwhile, since most of smartphones provide WiFi and Bluetooth communication services, indoor positioning services based on WiFi and BLE iBeacon (Bluetooth) can use smartphones as user premise equipment as well. For the case of RFID-tags, a RFID reader as user premise equipment is required. Some advanced smartphones support NFC (near-field communication) as RFID readers. However, some smartphones (e.g. iPhone) make some restrictions to apply NFC to third-party applications for security reasons. Regarding reachability, our proposed scheme uses pedestrian directional signs as the target objects to be included in the uploaded image. Since most of pedestrian directional signs are well installed to be seen as clearly as possible by visitors, it would not be hard for visitors to find any of them around their current location if they are inside the service area. For those using radio signals to transfer location information directly or indirectly, such as WiFi, Bluetooth, and Zigbee, users should be as close as to the range covered by the radio signal. In general, WiFi has better signal coverage than both Bluetooth and Zigbee on the basis of the same number of signal transceivers.

Aside from the approach based on active signal transceivers, another approach is based on passive sensor tags. Each sensor tag is assigned with a unique identifier. These sensor tags are then well deployed at different designate locations. The binding between sensor tags and their corresponding locations is preset. When a user is close enough to a sensor tag, by identifying the sensor tag, the user can get the location information associated with this sensor tag and thus get to know his/her current location. Among many others, RFID-tags and QR-codes are popular passive sensor tags for indoor positioning services. Compared with active sensor approaches, indoor positioning systems based on passive sensor tags can be built with less deployment cost. Some

passive sensor tags, for example RFID-tags and QR-codes, require tag readers keeping close enough to retrieve tag data. The good side is that passive sensor tags can provide high precision of location information if users get very close to the tag to get the information. However, the bad side is that the user should go close enough to get the tag data, which is not always an easy task to be done especially in a crowded area, such as the TMS case studied in this paper.

Our proposed scheme is based on identifying passive sensor tags as well. The proposed scheme estimates user location by identifying the pedestrian directional sign near the user. Compared with another two popular passive sensor-tag approaches, RFID-tags and QR-codes, the proposed approach has some additional benefits. First, the cost of deploying the passive sensor-tag is nearly free. We use available pedestrian directional signs in the service area as location landmarks without extra effort to setup the sensor tags. Second, most of pedestrian directional signs are well installed to be seen as clearly as possible by visitors, it would not be hard for visitors to find any of them around their current location if they are inside the service area. In contrast, how to deploy RFID-tags/QR-codes in a place where users can easy find them is not an easy task, as it is hard to find a small target such as RFID-tag/QR-code in a large and crowded area. Third, in a crowded area such as TMS, it would be easier to find a pedestrian directional sign and then take a picture of it, compared with the task of finding a RFID-tag/QR-code and then getting close to it to read the tag data. Last but not least, it can provide extra heading information besides the location information. Heading information is useful to help visitors to walk on the right direction toward their destinations. Since the orientation information of each pedestrian directional sign is known, once we identify the sign present in the uploaded image, we can provide the user with both of the location and orientation information about the sign.

Our proposed scheme uses signage as landmarks. We use TMS as an example to validate the feasibility of the proposed scheme. We believe that with careful design, similar system frameworks can be applied to many indoor environments such as an airport and a big museum, to name a few. Nonetheless, the proposed scheme has some limitations. First it cannot provide very precise location information. The resolution of location information depends on the deployment density of the pedestrian directional sign. Take the TMS case as an example. The distance between two neighboring signs is around 15–30 meters. While the geolocation resolution is not high, it is definitely good enough to help visitors to know where they are in the area. Second, it cannot be applied to all indoor scenarios but only those with well deployment of information signage for visitors. Meanwhile, another possible problem is due to the duplication of signage at different locations. In real-world, same signage could be installed in multiple different places. For such a case, our proposed system will inform the user existence of multiple same signs as the one taken by the user and show all locations of the signs in the map. And, then our proposed system would suggest the user to take another different signage nearby. According to our survey, among the 52 pedestrian directional signs we used in the TMS field experiments, there is only one sign which has been installed at multiple different places. In the field try area, there are two same pedestrian directional signs in a long corridor guiding visitors to the Taipei MRT Station having been installed on both sides of an escalator toward the platforms of the MRT station. Last but not least, the proposed scheme is designed mainly for indoor positioning services, which provides location and heading direction information to users according to their uploaded images. While both of the user location and heading direction information are useful for navigation service, there are some limitations to extend the proposed scheme for navigation services. For example, for a good navigation service, continuous

acquisition of positioning information is important to keep users all the time to stay on the right track. The requirement is subject to the density of the landmark (signage) deployment for our proposed scheme. If the signage installed in the service area cannot provide a network of seamless landmarks for users traveling from place to place in the server area, our proposed scheme cannot guarantee to provide the seamless positioning information to the user for indoor navigation services.

4. Pedestrian directional signage recognition

As shown in Figure 2, one core module of the proposed system framework is for detection and recognition of the pedestrian directional signs in the uploaded image. To this end, the Google Objection Detection framework is applied. In this section, we present design details of the signage recognition module. Taipei Main Station (TMS) is the largest train station in Taiwan. According to the number of annual visitors, it is one of busiest transportation station in the world as well. Consequently, TMS is an ideal target place for our experiment study for the proposed indoor positioning scheme. In the following subsections, we first introduce the pedestrian directional signs used in our experiments. Then, we present the details of the training image preparation and the bounding box schemes required in the model training procedure for the pedestrian directional signage image object detection.

4.1. Pedestrian directional signage in TMS

The B1 core area (Figure 1b excluding the park lot area) of TMS is the field try area, in which there are totally 52 different pedestrian directional signs included in this field try area. Figure 3 shows some examples of the directional signs in the field try area. Images 1–5 in Figure 3 are full-size picture images acquired by smartphones. To save the context page length, images 6–11 in Figure 3 just show the signage parts of the original picture images. For example, image 8 is taken from the picture shown in image 5 in Figure 3. In general, there are two types of pedestrian directional signs in the area. One is for general directional information (the one in yellow words) such as images 1–4 and 9–11 in Figure 3; the other is specially for traffic guidance toward MRT (Mass Rapid Transit) Taipei Station (the one in white words) such as images 5–8 in Figure 3. Most of the signs include text information (both in Chinese and English), icons, and arrow signs to indicate the direction. Many of the signs guide visitors to some hot spots (such as MRT/TR/HSR platforms, picket offices, main streets, restaurants, and facilities, to name a few) from different places. Consequently, some of them look very similar. For example, the signs in images 1–3 in Figure 3 have exact same context except the arrow signs and so do images 6–8 in Figure 3. All of the 11 images in Figure 3 are taken from 10 different signs hanging from the ceiling at different locations in the area (images 8 and 5 are from the same picture for the purpose of explanation).

Figure 3. Examples of pedestrian directional signage in Taipei Main Station (TMS).

4.2. Training image collection

For the signage image recognition experiments, we collect a lot of signage images of the 52 different pedestrian directional signs. We simulate possible user positions when they take pictures for the signs. For each of the signs, we take the signage image both from three different distances (Figure 4b) and 13 different viewing directions (Figure 4a) to the sign as shown in Figure 4. For each of the 52 pedestrian directional signs, first we stand straightly toward the sign (that is perpendicular to the sign), and then we find the nearest distance toward the signage to include the whole signage image into the field of view (FOV) of the smartphone camera. Then, from the same viewing direction, we walk three steps (about 1.5–2 m) away from the sign and take the second image. And then another 3-step distance are taken further away from the sign and the third picture of the same sign is taken. Similar procedures are applied to take another three images of the same sign from each of the 13 different viewing direction. As shown in Figure 4a, for each of the signs, we take the signage image from 13 different viewing directions, which are from 22.5 degrees to 157.5 degrees at an interval of 11.25 degrees approximately. Accordingly, thirty-nine ($3 \times 13 = 39$) images are taken for each sign from 39 different positions in each run. To enrich the training images, the same procedures have been done twice for each of the 52 pedestrian directional signs. That is, for each sign, we collected $39 \times 2 = 78$ images which roughly cover all of the combination of the three distances and the 13 viewing directions two times. Totally, we have $78 \times 52 = 4{,}056$ training images for the 52 signs used in the field try. All of the training images were taken by the same user with the

same smartphone (iPhone 6s) in two days at different time slots including day and night. Besides, all the training images were taken without room-in or room-out operations. The image size and the resolution were fixed with 1024 × 768 and 72 dpi respectively. Some of the training images are shown in Figure 3.

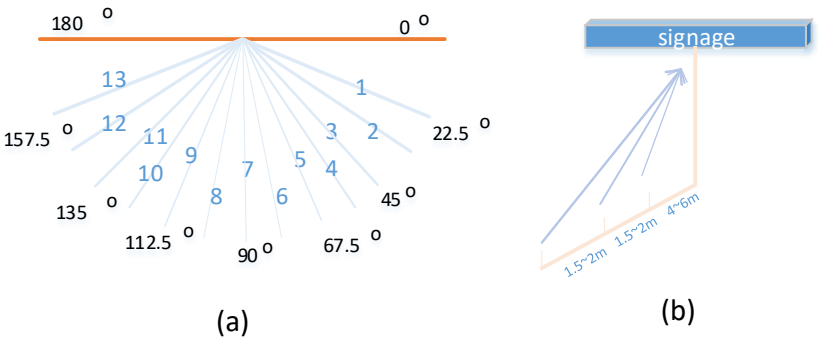

Figure 4. Training image collection scheme. For each signage, we take the signage images from different distances (b) and at different viewing directions (a).

4.3. Bounding box scheme and the image flipping option

We use the Google TensorFlow Object Detection framework for the image object detection and recognition. According to the framework, we need to label the target object to be detected with a bounding box for each of the training images. How to do the object tagging is beyond the scope of the object detection framework, and it should be subject to the applications. For our use cases, we need to identify the pedestrian directional sign in the uploaded image. In general, as shown in Figure 5, there are three different bounding-box schemes for the signage objects. The first one, denoted as *with-boundary (WB)*, is to include the whole sign including the boundary of the sign (Figure 5a). This scheme keeps most information of the sign, but it might include some background image context (that is those between the bounding box and the sign boundary) as well. To mitigate these extra background image context, the second bounding box scheme, denoted as *without-boundary type-A (WoB-A)*, is considered (Figure 5b). The *WoB-A* scheme excludes the sign boundary to remove the background image context as much as possible. Similar to the *WoB-A* scheme, the third scheme is to include the information context (such as texts, icons, and directional arrows) only, which is referred as *without-boundary type-B (WoB-B)* in this research study (Figure 5c). In the third scheme, we try to make the bounding box for a signage object as tight as possible. All the three bounding-box schemes are included in this research study. We would like to know if the three bounding-box schemes make any difference to the accuracy of the signage object detection. And, if any, then we would like to know what is the best bounding-box scheme for the signage object detection application. The answers to the above problems could help us to build a better signage object detection module for our proposed indoor positioning services.

Another training model design option we would like to investigate is about the image-flipping option. To enrich the training images, one common approach is to artificially generate a flipping image for each of the training image. For many natural objects, they preserve some symmetrical properties of their appearance. For example, we can (horizontally) flip a car image in which the car is

running from left to right to obtain a car image in which the car is running from right to left. This virtually generated car image would help the system to identify a car object in a test image which includes a car running from right to left. However, possible side-effects of enabling image-flipping are to increase possible confusion between the virtually flipped images and other images. For the case of signage object detection, since the main context in a pedestrian directional sign include text (both in Chinese and English), icons and arrow signs, it seems impossible to find a flipped sign in real life. Consequently, we suspect that enabling the flip-image option cannot improve accuracy of the signage object detection but increase confusion instead.

(a) bounding box scheme 1: with boundary (wb)

(b) bounding box scheme 2: without boundary type A (woba)

(b) bounding box scheme 3: without boundary type B (wobb)

Figure 5. Three different bounding-box schemes for the signage objects.

4.4. Training model design alternatives

How to design a signage object detector based on the Google Object Detection framework is a key design issue in this research study. There are design factors needed to be studied to produce a high accurate object detector. In this research study, we focus on three of those factors: the number of training images, bounding box schemes, and the image flipping option as we discussed in previous sections (Section 4.3). For the number of training images, we compared the accuracy of the detector trained on total training images with that trained on a half of the training images. For the full set of training images, each sign has two training images taken from each of the 39 positions as discussed in previous sections (Section 4.2). For the set including a half of the training images, each sign has one training image taken from each of the 39 positions. For the bounding-box schemes, the three different bounding-box schemes presented in Section 4.3 are considered. Last but not least, for the image flipping option, we compared the training model with and without enabling the image flipping option during model training.

5. Performance evaluations

5.1. Testing data sets

As shown in Table 2, we prepared three test data sets for performance evaluation. The first data set (dataSet1) has 2600 signage images, which were taken by the same smartphone (iPhone6s) used for collecting the training images. For each of the 52 signs, we took 50 images covering all the 39 positions designed for the training image collection process aforementioned. Since the test images in

dataSet1 were taken with the same smartphone on a different day and from positions covering those positions designed for collecting the training images, dataSet1 can be though as a baseline test image set to evaluate the signage detection accuracy. For collecting the second test image data set (dataSet2), besides the iPhone6s we included another two models of smartphones (HTC and Samsung). The second test data set was collected by three participants. Each participant used one of the three smartphones (iPhone6s, HTC, and Samsung). For each of the 52 signs, each participant individually collected 20 images from twenty different positions selected randomly among those 39 positions aforementioned for training image collection. Note that the position selection rule is just a rough guideline to each of the participants. Exact positions taken by each participant might not be the same as those 39 positions used by training image collection. One of main objectives to include dataSet2 is to enrich the variety of the test images, which are contributed by different participants with different models of smartphones and taken from different locations at different time slots on different days. For the third data set (dataSet3), we had another three participants to collect the test images with the same three smartphones as used for dataSet2. In order to make dataSet3 different from dataSet1 and dataSet2, in which we would like the participants to take the signage images at the locations similar to those used in the training image collection process, we let the participant to freely take the signage images for dataSet3 as a real user. The only guideline for them are to include the signage in the center of their smartphone screen.

Table 2. Test image data sets.

	dataSet1	dataSet2	dataSet3
smartphones	iPhone	iPhone/HTC/Samsung	iPhone/HTC/Samsung
Total images	2600	3120 (1040/1040/1040)	621(204/224/193)
#images per signage	50	60(20/20/20)	8-22(various)
Image collectors	One (same as the one prepare the training image)	Three (iPhone user is the same as the one in dataSet1)	Another three
Description	Same procedure as training image set	Random select 20 locations from the 39 candidate locations	Only general guide line is given, no special limitation

Table 3. Experimental results of signage object detection (with image-flipping On setting).

Bounding box sch.	With Boundary (*WB*)		Without Boundary-A (*WoB-A*)		Without Boundary-B (*WoB-B*)		Average	
accuracy	#errors/ total images	Correct rate	#errors/ total images	Correct rate	#errors/ total images	Correct rate	#errors/ total images	Correct rate
dataSet1	12/2600	0.9954	5/2600	0.9981	2/2600	0.9992	6.3/2600	0.9976
dataSet2	190/3120	0.9391	171/3120	0.9452	206/3120	0.9340	189/3120	0.9394
dataSet3	10/621	0.9839	19/621	0.9694	5/621	0.9919	11.3/621	0.9818
Overall (dataset 1+2+3)	212/6341	0.9666	195/6341	0.9692	213/6341	0.9664	206.7/6341	0.9674

Figure 6. Confusion matrices of the signage object detection results (with image flip-On setting). The rows from top to bottom are for dataSet1 (DS1), dataSet2 (DS2), and dataSet3 (DS3) respectively; the columns from left to right are for the three bounding boxes: with-boundary (*WB*), without-boundary type-A (*WoB-A*), and without-boundary type-B (*WoB-B*).

5.2. Testing results under image flip-on training model

First, we evaluate the accuracy of the signage object detection with enabling the image flipping-option. The testing results for all the three test data sets are shown in Table 3. In the signage image object detection experiments, there are 52 different signs to be detected. For each of the test images, the object detector built based on the Google Object Detection framework is used to detect possible signage objects in the test images. Every detected signage object is associated with a confidence value. If multiple signage objects are detected, we take the one with the highest confidence value as the identified pedestrian directional sign. Figure 6 shows the confusion matrix of the experiment results. The confusion matrix is a 52×53 matrix. The last column (column 53) is for "no signage" case, which is corresponding to the result that the object detector did not find any signage object in the test image. According to our experiment results, we found about near half of the

misclassification cases are belong this type of errors.

The experimental results shown in Table 3 indicate that the signage object detector we constructed performs very well. The overall average correct rate is 96.74% (= 1−206.7/6341) regarding all the 6341 test images contributed from the three data sets and the average number of errors under the three different bounding box schemes. Regarding individual test data sets, the test results of the dataSet1 demonstrate great accuracy (99.76%) of signage object detection. There are, in average, only 6.3 images being detected incorrectly over the 2600 test images. The test results from dataSet2 degrade the object detection accuracy with a few percentages to 93.94% in average, compared with the results for dataSet1. The test images of the two data sets are collected by different participants on different days with different models of smartphones. For dataSet2, we included two another models of smartphones (HTC and Samsung) to take the test images, besides the one (iPhone6s) used in the training data collection. However, we do not think that the degradation of detection accuracy is mainly caused by the use of different models of smartphones. The test results of dataSet3 show that with the same three smartphones, the detection accuracy is also as high as 98.18% in average. We have reviewed those test images being identified incorrectly. Over half of them are due to poor signage image quality. They are either too small (the user is too far away the sign) or the shooting angle is too oblique to the sign. We show some of incorrect classification examples in Figure 7. Comparing those training image examples as shown in Figure 3, the signage objects shown in these misclassified test images are apparently small than those in the training images.

Meanwhile, we found that the three different bounding box schemes did not make much difference for the object detection accuracy. The bounding box scheme with the *WoB-B* strategy performs slightly better than the other two in the test data set dataSet1 and dataSet3, but it is the worst one among the three regarding the test data set dataSet2. The best bounding box strategy for the test data set dataSet2 is with the *WoB-A* strategy, while it is the worst strategy for the test data set dataSet3.

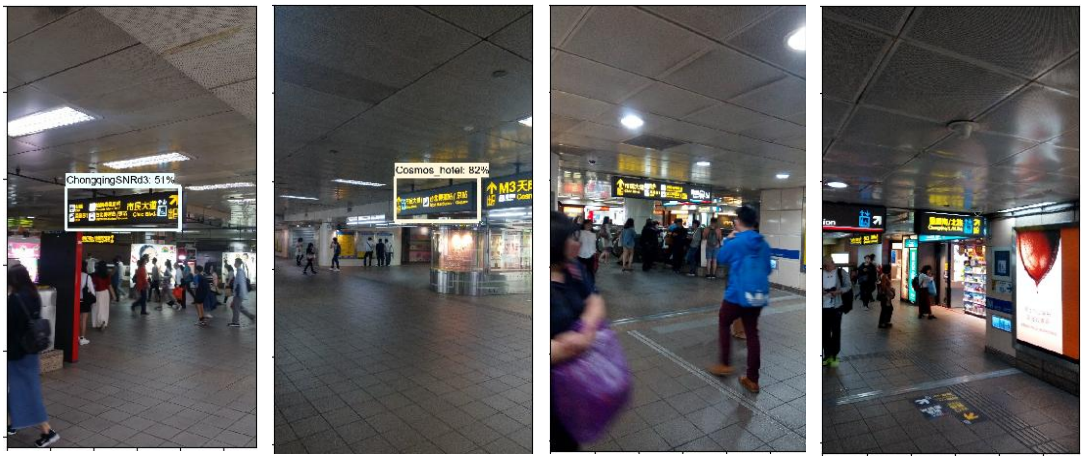

Figure 7. Examples of the failed test images. The first two images are examples of misclassification error (signage objects have been detected but misidentified); the other two are example of misdetection error (no signage object has been detected).

5.3. Testing results under image flip-off training model

To enrich the training images, one common approach is to artificially generate a flipping image for each of the training image. For many natural objects, they preserve a symmetrical property of their appearance. However, for the case of signage object detection, since the main context in a pedestrian directional sign includes text (both in Chinese and English), icons and arrow signs, it seems impossible to find a flipped pedestrian directional sign in real life. Consequently, we suspect that enabling the flip-image option might not improve accuracy of signage object detection but increase confusion instead. To investigate this conjecture, we turned off the image flipping option of the training model and retrained the object detection network with same settings as before. All the model training parameters were set to be the same except the image-flipping option involved in the model training. Then, we evaluated the new object detection network with all the test images in the three data sets. The experimental results are shown in Table 4.

Table 4. Experimental results of signage object detection (with image-flipping Off setting).

Bounding box sch.	With Boundary (*WB*)		Without Boundary-A (*WoB-A*)		Without Boundary-B (*WoB-B*)		Average	
accuracy	#errors/ total images	Correct rate	#errors/ total images	Correct rate	#errors/ total images	Correct rate	#errors/ total images	Correct rate
dataSet1	4/2600	0.9985	6/2600	0.9980	5/2600	0.9981	5/2600	0.9981
dataSet2	117/3120	0.9625	111/3120	0.9644	128/3120	0.9590	118.7/3120	0.9620
dataSet3	8/621	0.9871	4/621	0.9936	10/621	0.9839	7.3/621	0.9882
Overall (dataset 1+2+3)	129/6341	0.9797	121/6341	0.9809	143/6341	0.9774	131/6341	0.9793

As shown in Table 4, the overall average correct rate is approaching 98% (= 1−131/6341). In average, among all the 6341 test images included in the all three test data sets, there are only 131 images are misclassification or misdetection. Compared with the results for the cases of enabling the image-flipping option (Table 3), the average number of incorrect test images is reduced from 206.7 to 131 over the 6341 test images, which is a nontrivial improvement. Regarding the improvements on individual test data sets, the results show that the average detection accuracy over the three bounding box schemes is improved as well for all the three test data sets, especially for dataSet2.

As for the differences of detection accuracy under the three bounding box schemes, again we found that three is no significant difference between the three different bounding box schemes regarding the object detection accuracy, under the new setting (disable the image-flipping option). Different from the cases of enabling image-flipping option, the bounding box scheme with the *WoB-A* strategy performs slightly better than the other two in the test data set dataSet2 and dataSet3, and it performs nearly well in dataSet1. Regarding the total number of error detection counts over all the three test data sets, the *WoB-A* bounding box scheme slightly outperforms the other two, which is the same as for the cases of enabling the image-flipping option as shown in Table 3.

5.4. Performance impacts of training data size

Last but not least, we would like to know the performance impacts of training data size on the

proposed scheme. In general, larger training data size is beneficial to generate a more accurate object detection network, but it, on the other hand, increases the cost of training data collection and the time required for the network model training. In the original setting, we use all the training images we collected, in which for each sign we collect 78 test images equally distributed from the 39 reference locations as presented in Section 4.2. We then reduce the training data size to one half, in which for each signage we take one image at each of the 39 reference locations. The evaluation results are shown in Table 5. Two data sets, dataSet1 and dataSet3, are used for the performance comparison. In Table 5, the row labeled as "−78" is for the cases with full training data (78 images per signage), and the row labeled as "−39" is for the cases with half training data (39 images per signage). The results show the average accuracy (correct rate) drops from 0.9976 (with full training data size) to 0.9821 (with half training data size) for test data set dataSet1. As for the case of test data set dataSet3, the average accuracy drops from 0.9818 to 0.9436. While we can have a fair accuracy (over 94% in both dataSet1 and dataSet2 test data sets) with only 39 training images for each sign, it seems still worthy to achieve over 98% of accuracy at the cost of doubling the training image size (78 training images for each signage). According to our experience, it takes less than half of a day with one manpower to collect the training images. And the training process for the object detection network can be done in less than two days in our experiments.

Table 5. Performance impacts of training data size (with image-flipping On setting).

Bounding box sch.	With Boundary (*WB*)		Without Boundary-A (*WoB-A*)		Without Boundary-B (*WoB-B*)		Average	
accuracy	#errors/ total images	Correct rate	#errors/ total images	Correct rate	#errors/ total images	Correct rate	#errors/ total images	Correct rate
dataSet1								
−78	12/2600	0.9954	5/2600	0.9981	2/2600	0.9992	6.3/2600	0.9976
−39	44/2600	0.9831	63/2600	0.9758	33/2600	0.9873	46.7/2600	0.9821
dataSet3								
−78	10/621	0.9839	19/621	0.9694	5/621	0.9919	11.3/621	0.9818
−39	46/621	0.9259	36/621	0.9420	23/621	0.9629	35/621	0.9436

6. Conclusion

In this paper we investigate the feasibility of an infrastructure-free intelligent indoor positioning system based on visual information only. The proposed scheme is different form the conventional infrastructure-based approach, in which service providers usually need to deploy additional sensor devices to the environment, and require elaborate sensor deployment topology design and calibration procedures. To eliminate these extra cost and efforts, the proposed scheme utilizes available objects in the environment, such as pedestrian directional signs in a transportation station, as location landmarks. Leveraging wildly available smartphone devices as customer premises equipment to the users and the cutting-edge deep-learning technology, we have demonstrated the proposed scheme is practical and feasible to be deployed in a public service area by a field try in Taipei Main Station.

In the proposed scheme, we use available pedestrian directional signage as location landmarks. Via a pre-download smartphone App, a user can get his/her current location by uploading an image of a nearby pedestrian directional sign to the positioning system. Then, the system identifies the pedestrian directional sign in the uploaded image and thus indicates the current location of the user.

In our field try experiment, 52 pedestrian directional signs are included to be identified in the testing area. The content of each sign consists of three parts in general: the street/building/facility name both in Chinese and English, and the icon for direction sign and the facility. The Google Objection Detection framework is applied for signage detection and recognition. According to the experimental results, we have shown that the proposed system can achieve as high as 98% accuracy to correctly identify the pedestrian directional sign in the testing image over the 6341 test images.

Some of the key design factors of the signage image object detection are discussed. Regarding the training data size, from the experimental results, we have found that in the proposed object detection framework, 78 training images per pedestrian directional sign is sufficient to provide as high as 98% accuracy of signage object identification among the 52 possible candidates. Meanwhile, we have also found that disabling the default setting of the image-flipping option can increase non-trivial identification accuracy in all of the three test data sets. It can significantly reduce one-third of the average number of misclassification/misdetection test images from 206.7 to 131 among all of the 6341 test images. Last but not least, for the bounding box schemes, the experimental results indicate that the three possible schemes, *with-boundary* (*WB*), *without-boundary type-A* (*WoB-A*), *and without-boundary type-B* (*WoB-B*), have no significant difference in the detection accuracy. They all can achieve more than 97.7% of detection accuracy regarding all the 6341 test images in the three test data sets under the image flipping-off setting. Nonetheless, the *WoB-A* outperforms the other two slightly, with around 98.1% of detection accuracy over all of the test images.

Acknowledgments

This work was supported, in part, by Ministry of Science and Technology, Taiwan, under Grant No. MOST107-2221-E019-044.

Conflict of interest

The authors declare that there is no conflict of interests in this paper.

References

1. K. Wang, X. Yu, Q. Xiong, et al., Learning to improve WLAN indoor positioning accuracy based on DBSCAN-KRF algorithm from RSS fingerprint data, *IEEE Access*, **7** (2019), 72308–72315.
2. L. Chen, B. Li, K. Zhao, et al., An improved algorithm to generate a Wi-Fi fingerprint database for indoor positioning, *Sensors*, **13** (2013), 11085–11096.
3. C. H. Lin, L. H. Chen, H. K. Wu, et al., An indoor positioning algorithm based on fingerprint and mobility prediction in RSS fluctuation-prone WLANs, *IEEE T. Syst. Man. Cy. S,* **6** (2019), 1–11 (Early Access).
4. Md. S. Ifthekhar, N. Saha and Y. M. Jang, Neural network based indoor positioning technique in optical camera communication system, in Proceedings of 2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN), *IEEE Comput. Soc.*, (2014), 431–435.

5. G. Schroeer, A real-time UWB multi-channel indoor positioning system for industrial scenarios, in Proceedings of 2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN), *IEEE Comput. Soc.*, (2018), 1–5.
6. X. Gan, B. Yu, Z. Heng, et al., Indoor combination positioning technology of Pseudolites and PDR, in Proceedings of 2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS), *IEEE Comput. Soc.*, (2018), 1–7.
7. J. Huang, V. Rathod, C. Sun, et al., Speed/accuracy trade-os for modern convolutional object detectors, in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), *IEEE Comput. Soc.*, (2017), 3296–3297.
8. R. Girshick, J. Donahue, T. Darrell, et al., Rich feature hierarchies for accu-rate object detection and semantic segmentation, in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), *IEEE Comput. Soc.*, (2014), 580–587.
9. S. Ren, K. He, R. Girshick, et al., Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE T. Pattern Anal.*, **39** (2017), 1137–1149.
10. J. Dai, Y. Li, K. He, et al., R-FCN: object detection via region-based fully convolutional networks, preprint, *CoRR*(2016), arXiv:1605.06409.
11. J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, et al., Selective search for object recognition, *Int. J. Comput. Vision*, **104** (2013), 154–171.
12. J. Redmon, S. K. Divvala, R. B. Girshick, et al., You only look once: Unified, real-time object detection, *CoRR*(2015), arXiv:1506.02640.
13. W. Liu, D. Anguelov, D. Erhan, et al., SSD: Single shot multibox detector, in Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision–ECCV 2016. *Lect. Notes Comput. Sc.*, **9905** (2016), 21–37.
14. T. Lin, P. Goyal, R. Girshick, et al., Focal loss for dense object detection, in Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), *IEEE Comput. Soc.*, (2017), 2999–3007.
15. YOLO: Real-Time Object Detection. Available from: https://pjreddie.com/darknet/yolo/.
16. GitHub, tensorflow/models. Available from: https://github.com/tensorflow/models/tree/master/research/object_detection.
17. Taipei Main Station, Wikipedia. Available from: https://en.wikipedia.org/wiki/Taipei_Main_Station.
18. Statistics inquiry, Ministry of transportation and Communications, Taiwan. Available from: http://stat.motc.gov.tw/mocdb/stmain.jsp?sys=100&funid=emenu.
19. London Waterloo station, Wikipedia, Available from: https://en.wikipedia.org/wiki/London_Waterloo_station.

AIMS Press