Universal Access in the Information Society (2023) 22:69–82
https://doi.org/10.1007/s10209-021-00837-9

LONG PAPER

A dataset for the recognition of obstacles on blind sidewalk

Wu Tang[1] · De-er Liu[1] · Xiaoli Zhao[2] · Zenghui Chen[1] · Chen Zhao[3]

Accepted: 4 August 2021 / Published online: 16 August 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Recently, the technology of assisting the navigation of visually impaired persons with computer vision has been greatly developed. A number of scholars have conducted related research, including indoor and outdoor object detection for blind people. However, there are still problems with some existing methods or datasets. Our work mainly proposes a dataset (OD) for assisting the detection and recognition of outdoor obstacles for blind people on blind sidewalk. We classify some common obstacles, train the dataset with state-of-the-art detectors to obtain detection models, and then analyze and compare these models in detail. The results show that our proposed dataset is very challenging. The OD and the detection model can be obtained at the following URL: https://github.com/TW0521/Obstacle-Dataset.git.

Keywords Visually impaired person (VIP) · Obstacle detection · Blind sidewalk · OD

1 Introduction

In this era of science and technology leading mankind to advance, computers have become an indispensable tool, and the development of computer vision has brought tremendous convenience for human progress. At present, computer vision has been widely substituted for face recognition [1–3], text recognition [4–6], public safety [7, 8], industrial applications [9] and other fields [10–12]. It can be said that computer vision is closely related to our lives, which is equivalent to the "third eye" of human beings. However, most of the current applications are studied based on normal human vision. As the world population grows, the number of visually impaired persons (VIP) also increases. Researches have shown that individual blind person' difficulties in getting around in outdoor environments mainly include finding a certain place, crossing the street, etc., where more detailed problems include the inability to read street signs and traffic lights. They believe that the external environment has the greatest impact because there are more obstacles around the blind sidewalk. They expected fewer obstacles around the sidewalks and to be wider again [13]. Outdoor traveling is a major difficulty for blind people, who may face a variety of obstacles, and identification of these obstacles can effectively help them perceive their surroundings. As society pays close attention to them, the issue of assisted blind travel has received considerable critical attention.

Scholars have begun to research in this area and have successively introduced some equipment and systems for visually impaired people to detect obstacles when traveling.

KR-VISION Technology Co., Ltd [14] has developed a set of lightweight obstacle avoidance glasses (Intoer) for the blind that includes three main functions: obstacle and pathway detection, scene detection, and precise location navigation. This requires a large obstacle dataset and detection model to support the functionality of the glasses. It is also demonstrating the flexibility of computer vision-assisted devices. In order to help blind people to travel better, Tapu et al. [15] used multi-scale algorithms to extract points of interest from the video and use them to detect obstacles. After detecting the obstacle, the obstacle is marked as emergency or normal according to its motion status. However, this method cannot detect multiple objects at the same time. In [16], the authors labeled obstacles and captured them with a camera, and then used OpenCV to separate and recognize the labeled images. This method requires tagging

✉ De-er Liu
landserver@163.com

[1] School of Civil and Surveying & Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, Jiangxi, China

[2] School of Economics and Management, Jiangxi University of Science and Technology, Ganzhou 341000, Jiangxi, China

[3] Fujian Jingwei surveying and mapping information CO., Ltd, Fuzhou 350000, Fujian, China

of obstacles and needs high labor cost. In [17], the authors proposed an assistive device for visually impaired people, which can help them to complete path navigation and process the collected images to detect obstacles, using YOLO v3 detector to train the detection model, which is able to recognize 6 kinds of objects (person, pit hole, car, stairs, chair, washrooms), and its detection accuracy reached 94%. Although its detection accuracy is better, it can recognize fewer categories and cannot meet the needs of blind people to travel. In [18], the authors proposed an unknown environment obstacle classification method based on smartphones and computer vision, which was able to classify objects into three major categories: vehicle, person and others. The method achieved a high accuracy through experimental evaluation. However, this method only classifies better in the above three categories and lacks the refinement of obstacle categories. For indoor environments, [19] proposed an indoor sign and door detection system, which is built using a ResNet and is capable of identifying four types of sign information in the room. Besides, [20] also proposed an indoor object detection scheme for indoor activities of blind people, which was built based on the RetinaNet framework and evaluated using different backbones such as ResNet (residual network) and DenseNet. Moreover, they proposed an Indoor Object Detection and Recognition Dataset (IODR, Indoor object detection and recognition dataset), which contains 14 categories, where the training set contains 5300 images and the test set contains 2700 images. For blind outdoor travelers, [21] used Google target detection API to detect and identify crosswalk lights and bollards, and provided real-time feedback to blind users on the detection results. The detection system was evaluated using SSD and RCNN, and good detection results were achieved. Similarly, [22] developed an outdoor recognition system that recognizes the color of crosswalk light, bollards and Braille blocks. The authors used a programmable remote control vehicle to evaluate the detection system concluding that the method is effective in guiding blind people to avoid obstacles. The research of many scholars mentioned above has promoted the development of computer vision to assist the visually impaired and also expanded the application field of computer vision. Many results achieved high detection accuracy but were able to detect fewer categories and lacked suitable benchmark datasets for obstacle detection. More accurate models and powerful datasets are needed provide support. Therefore, our main contributions to this work are described in the following:

Obstacles were systematically classified, and a dataset OD (obstacle dataset) was created to support travel obstacle detection for the visually impaired, which can be used to develop and evaluate obstacle detectors.

We performed benchmark tests on the obstacle dataset using a variety of state-of-the-art target detectors and developed a basic model for obstacle detection.

2 Motivation

In the field of object detection datasets, Pascal VOC [23] (Pattern analysis, statistical modeling and computational learning Visual Object Classes), Microsoft COCO [24] (Common Objects in Context) and other general datasets occupy an unshakable position. They can be used not only as a dataset for object detection but also for semantic segmentation and instance segmentation. Pascal VOC contains VOC2005, VOC2006, VOC2007 and VOC2012. There are only five categories included in VOC2005, and VOC2006 is expanded to 10 categories. VOC2007 and VOC2012 expanded to 20 categories. The categories mainly included some objects and animals that are common in life, such as cars, cats and dogs. Compared to the COCO dataset, the VOC has a smaller number of pictures. The COCO dataset contains 91 object categories that are common in life, such as snowboards, fruits and transportation facilities. The COCO2017 contains a total of nearly 120 k images of these categories. The current state-of-the-art object detectors will use the above two datasets to train and evaluate detection models. In addition to these datasets, there are also pedestrian detection datasets [25–27], text detection datasets [28, 29], traffic signs detection datasets [30, 31], face detection datasets [32, 33], remote sensing image object detection datasets [34, 35] and others. These datasets are labeled and annotated with different methods and can be used for corresponding object detection in each adaptive scene. There are many other datasets of the same type in addition to those listed in the table, which will not be repeated here. With the development of deep learning and supervised learning, more scholars will conduct research on unsupervised and semi-supervised learning; however, at present, if you need to train a good model, you still need a lot of sample data.

Unlike conventional object detection, obstacle detection for visually impaired people traveling outdoors does not require too many useless categories, but only needs to classify common obstacles on the blind sidewalk. The above datasets have their application fields for different scenarios, and some of these categories are not considered obstacles within the scope of this research. What we need are some common obstacles that can really affect the travel of visually impaired people. For example, some categories

of fruits, food and tools in the COCO dataset cannot be seen in outdoor scenes. Therefore, if one wishes to detect and recognize certain types of obstacles, it is necessary to first establish a dataset of obstacles. In addition, to make the dataset reach its generalization and robustness, it is necessary to use some detectors to train the dataset. The collected pictures also need to be diversified, such as background, lighting, size and target angle. It is also necessary to divide the dataset to ensure that the training and evaluation of the data are relatively independent.

3 Image classification and annotation

3.1 Obstacle classification

The category of the obstacle is a difficult problem to determine as it has not been possible to identify some objects that hinder the progress of the blind people when they travel. In major cities in various countries of the world, there are blind sidewalks built to facilitate the travel of blind people. However, according to our survey, although many cities have built blind sidewalks, various vehicles and objects are parked on them. Secondly, in order to regulate the path of the blind sidewalk and ensure its smoothness, barricades have been erected around blind alleys to block access to large vehicles. These are potential obstacles that may hinder the blind person's advancement. As shown in Fig. 1a, b, the red line area is a special road for the blind to walk in the city. In order to allow the blind to maintain the correctness of the walking direction, different floor tiles are used for paving. Figure 1a shows that two different vehicles are parked on the blind sidewalk, and a pedestrian is approaching in the forward direction. The pedestrian can take the initiative to avoid it, but the vehicle cannot. Assuming that the blind is walking on this road, he/she will easily hit these vehicles without relying on any auxiliary equipment. If they deviate from the sidewalk, there will be a pole for lighting the streetlight around the sidewalk. As shown in Fig. 1b, four spherical barricades were placed at the intersection of the blind sidewalk and the intersection to prevent the entry of some large vehicles or equipment, but these can also become obstacles that prevent blind people from walking on this blind sidewalk. It is these immovable obstacles that pose the biggest challenge for visually impaired travelers.

There are also many objects that can be considered as obstacles on the blind path, and we have summarized some common obstacles by reviewing literature and surveys. These objects are classified according to their properties, as shown in Fig. 2.

We divide the obstacles on the blind sidewalk into two categories: dynamic and non-dynamic. Non-dynamic obstacles need to be actively avoided, while dynamic obstacles can be avoided without active avoidance. We use red, orange, blue and green to indicate these obstacles according to the degree of danger for blind. According to the classification system, we selected 15 specific objects as the constituent categories of the OD. Among them, the person means pedestrians and passersby; they can take the initiative to avoid, which is a dynamic obstacle. Pole represents some standing poles around the blind sidewalk, including telephone poles, communication poles, etc. If the blind person deviates from the blind sidewalk, then there is a probability of being blocked by it. The pole is a static permanent obstacle. Ashcan means trash bins, trash cans, etc. Ashcan may be located on the blind sidewalk or may be placed around there, including some permanently fixed trash bins and some temporary in a certain position. Motorbike, bicycle and tricycle indicate that some motorcycles, electric motorcycles, bicycles and tricycles parked around the blind sidewalk, these are passive avoidance obstacles. The dog may be a pet dog carried by some pedestrians or a guide dog that some visually impaired people need to travel, which is a dynamic active avoidance obstacle. The warning column and spherical roadblock are usually located at the intersection of blind sidewalk, which are used to regulate the passage of pedestrians and vehicles. They are fixed permanent obstacles. Although cars, trucks and buses are rarely encountered on blind sidewalk, they are generally located where blind sidewalk and

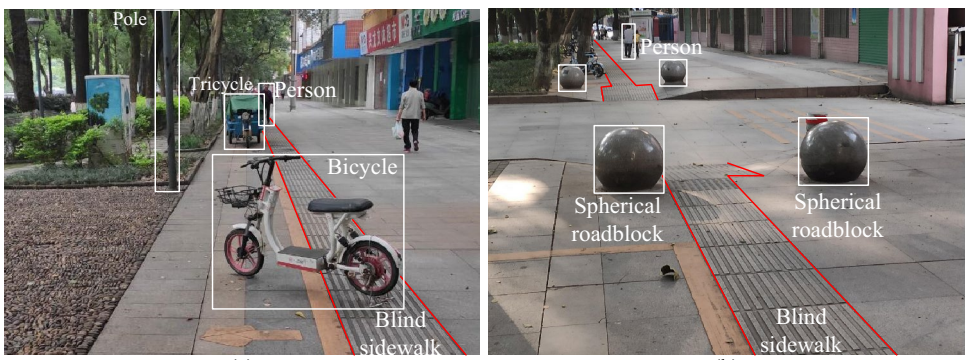

Fig.1 Potential obstacles on the blind sidewalk: **a** objects in and around the blind sidewalk, **b** spherical obstacles around the blind sidewalk to restrict vehicle entry

intersections cross. The stop sign is usually standing at a certain intersection. If blind persons encounter this, they have reached the intersection and need to be careful. Fire hydrant usually means some fire extinguishing equipment located around the blind sidewalk, which is a permanent facility. Reflective cones are usually located in some construction sites or used to guide pedestrians and vehicles, which are non-dynamic temporary obstacles. According to our survey, the above-mentioned 15 kinds of obstacles are often appearing around blind sidewalk.

3.2 Collection

With the obstacle categories identified, we began collecting these images. Regarding the collection of images, we determined several sources and first compared the categories based on the identified categories with those of the COCO dataset and the VOC dataset, from which we selected some categories of images and extracted the annotation information, include person, bicycle, car, bus, motorbike, dog, truck, stop sign and fire hydrant. However, these are only a part of the OD, and the remaining several categories are collected by using mobile devices for shooting and web. Frame shooting is required for video captured with mobile devices. Secondly, we also extracted some pictures from the TT100K (Tsinghua-Tencent 100 K) dataset. TT100K contains some pictures of intersections and traffic conditions, so it is more suitable for the dataset we established. The image resolution in the TT100K is relatively large, which can enhance the generalization performance of the dataset. However, because the TT100K labeling method is different from VOC, we extracted only part of the images and then annotated them.

3.3 Annotation

Currently, in the field of object detection, bounding box-based annotation methods are more common. The bounding box usually contains (xc, yc, w, h), where (xc, yc) are the center position, and w and h are the width and height of the bounding box. The format of the VOC dataset is mainly composed of images and annotations. Images are pictures used for training and testing, and annotations are saved annotation files corresponding to the picture, which is usually an xml file. The annotation files contain information such as the name of the labeled object, the size of the image and the coordinate position of the labeled frame. Therefore, we use the labeling method of the VOC and use the open-source labeling tool LabelImg [36] to label the collected pictures and generate the xml file.

The accuracy of the bounding box has a greater impact on model training. Therefore, in order to obtain high-precision object bounding boxes for pictures collected with mobile devices, we use manual methods to label instead of using automatic labeling. In order to normalize the obstacle categories, for images extracted from other datasets, we also screen and "re-label" obstacle objects in the images. For example, there is no pole objects in the categories extracted in the COCO,

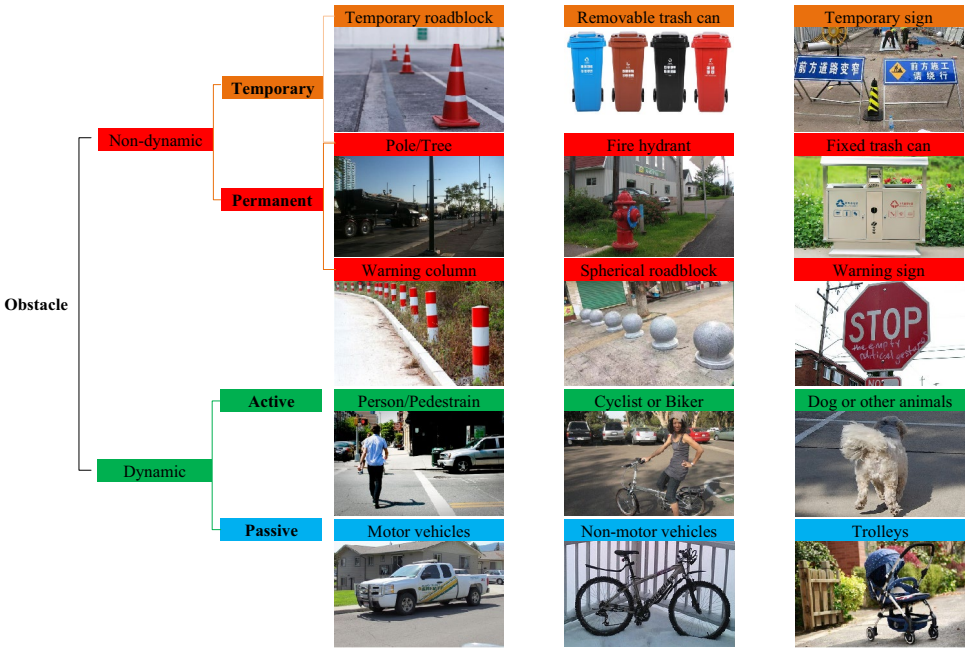

Fig.2 Obstacle division; the obstacles are divided into dynamic and non-dynamic obstacles according to the target attributes

(a) (b)

Fig.3 "Re-labeling" the picture: **a** the bounding box before the labeling, **b** the bounding box after the labeling

so it is necessary to check and label these categories that are not already in COCO, as shown in Fig. 3. Figure 3a shows the annotation information of the picture from the VOC. Before the re-labeling, only cars were labeled in this image, but we can see that the image also contains other objects we need. So, we "re-labeling" and draw a bounding box around the reflective cone and save it. We must check the images from other datasets to ensure that these objects in the images are labeled.

4 Properties of dataset

4.1 Number and division of OD

The number of images in the dataset directly determines the size of the data, which directly affects the effectiveness of subsequent experiments. Up to now, we have built a dataset containing a total of 7915 images with more than 40 k bounding boxes, 5066 images in the training set, 1583 images in the test set and 1266 images in the validation set. The specific quantity information is shown in Table 1.

The table shows the number of pictures and bounding boxes on the training set, test set and validation set for each category. The picture represents the number of images, BBox (bounding box) represents the number of bounding boxes, Total represents the total number, Train represents the training set, Test represents the test set, and Val represents the validation set.

4.2 Resolution analysis

The resolution of the image determines the quality of the image. In the dataset, image resolution has a greater impact on the training model. The general object detection model will normalize the size of the input image. For example, SSD (single-shot multi-box detector) will normalize the input image size to 300×300 when training the model. If the size of the input image is just 300×300, the features of the objects in the image will not be compressed. Conversely,

Table 1 Dataset categories

Dataset	Total		Train		Test		Val	
Name	Picture	BBox	Picture	BBox	Picture	BBox	Picture	BBox
Person	3228	13,828	2064	8813	647	2796	517	2219
Bicycle	746	1378	472	859	141	256	133	263
Car	2948	11,348	1092	7214	562	2108	484	2026
Motorbike	752	1671	486	1085	159	363	107	223
Bus	718	1059	456	676	135	184	127	199
Reflective cone	784	2820	495	1730	154	628	135	462
Truck	1560	2479	994	1594	288	455	278	430
Warning column	635	2356	415	1501	137	540	83	315
Dog	721	956	471	623	150	198	100	135
Spherical roadblock	462	1798	306	1204	82	314	74	280
Fire hydrant	938	1020	606	654	213	241	119	125
Stop sign	1091	1265	692	807	214	241	185	217
Ashcan	1073	1554	681	985	200	286	192	283
Tricycle	888	1028	572	667	180	207	136	154
Pole	2436	4283	1547	2726	493	896	396	661
Total	7915	48,843	5066	31,138	1583	9713	1266	7992

if the resolution of the input image is larger than 300×300, then the features in the image are compressed at the time of input with the image size. For a deep learning model, it needs to learn different size features of the same object. It needs to use images of different resolutions for training, so that the model can learn features on images of different resolutions, which can satisfy the requirements of model robustness and generalization. During training, we need to perform data enhancement operations, including image rotation, mirroring, etc. These algorithms generally provide image enhancement operations during training and testing.

In OD, we collected images of various sizes for training and testing. The resolution analysis is shown in Fig. 4. As can be seen (Fig. 4), regardless of the training set, test set or validation set, the resolution of most images is distributed within 1500×1500, and a few pictures exceed 3000×3000. We found that the resolution of the images extracted from the COCO dataset and the VOC dataset is basically within 1000×1000 pixels by analysis, while the resolution of the pictures taken through the mobile device is relatively large. The resolution of the images in the TT100K dataset is 2048×2048 pixels. The resolution distribution of the images is more diffuse, which plays a beneficial role in the model training dataset.

5 Experiments and evaluation

5.1 Object Detector for evaluation

In order to evaluate the OD and build an obstacle detection model, we used several different object detection models to train the dataset, including the typical two-stage detector Faster RCNN and single-stage detector SSD, YOLO (You Only Look Once) v3, YOLO v5 [37].

Faster RCNN [38] (faster region convolutional neural network) is a typical two-stage detector, and its detection stage are mainly divided into two parts. First, the detection frame is generated by RPN (region proposal network) and then sent to the classifier for classification and detection. Although the detection speed and accuracy of Faster RCNN in the detector at the same stage are improved, it is still far from the single-stage detector.

SSD was proposed by Liu [39] in 2015. SSD adopts the method of deriving multi-level feature layers for object detection. The proposal of SSD solves the problem of small object detection at that time.

YOLO [40] proposed to solve the problem of slow detection speed. The fastest version can even reach 155FPS, which maximizes the detection speed while ensuring the accuracy. However, in 2020, Joseph Redmon, the author of YOLO, publicly announced his withdrawal from the CV (Computer Vision) field. As of the time of withdrawal, Joseph Redmon has updated 3 versions of YOLO. The above detectors have become the current state-of-the-art object

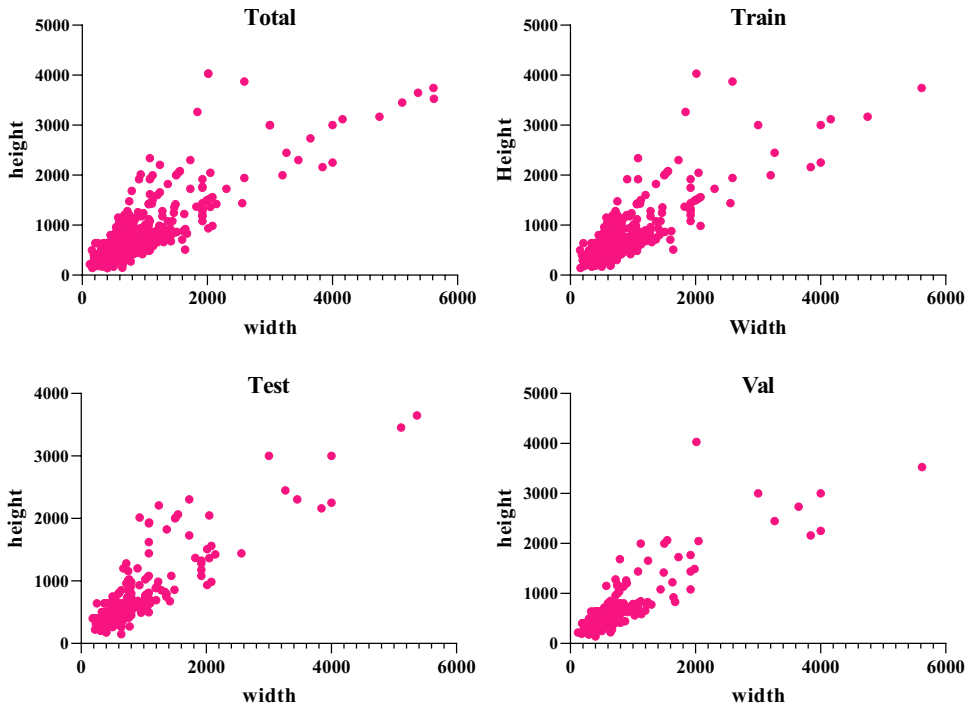

Fig.4 Resolution analysis of OD: **a** resolution of all images, **b** training set, **c** test set, **d** validation set

detectors, and some scholars have made improvements and applications around them [41–45].

5.2 Evaluation and analysis

In 2020, the v4 and v5 versions of YOLO were successively launched, and YOLO v4 was released by Alexey [46] in early 2020. Alexey's works has been affirmed by Joseph Redmon, and v3 has been upgraded and improved. Shortly after the release of v4, another version improved according to v3 (we called it v5 [37]) was released too. At this point, YOLO's big family has grown to v5. The experimental environment configuration is as follows: GTX2080TI, AMD Ryzen 9 3900X and 32 GB RAM. The experiment mainly uses the Please add section title: References at this point (v1.0) to train the OD. In order to contrast with the YOLO v5, we also used the SSD, YOLO v3 and Faster RCNN models to train the OD in the same way.

YOLO v5 contains four pre-trained models: YOLO v5s, YOLO v5m, YOLO v5x and YOLO v5l. We use these pre-trained models to train the OD and then use the test set and the validation set to evaluate the model. Using the v5 training dataset will save two model files, i.e., best and last. Therefore, after the training is completed, a total of 8 model files are included. The reason why YOLO v5 has four pre-trained models is that there are differences between the five models, some can achieve good accuracy but the model file is large, some models are small but the detection accuracy is slightly lacking, where the training speed will also vary depending on the pre-trained models. Figure 5 shows the results of the YOLO v5 trained OD model evaluated on the test set and validation set. There are four main indicators in total, where mAP@0.5 (mean average precision) means mAP when the threshold is set to 0.5, and mAP@0.5:0.95: means mAP with threshold is 0.7. Precision means detection accuracy, and recall means the recall rate.

Figure 5 shows that the highest mAP is the model pre-trained by YOLO v5x. mAP@0.5 of the trained YOLO v5x-last on the test set and the validation set are 0.761 and 0.765, respectively. The two values in the test and validation of mAP@0.5:0.95: are 0.55 and 0.547, respectively, and its recall and precision are also among the highest. However, the size of its model has reached 729 MB, which undoubtedly increases the calculation burden during detection. Among the four models trained, the smallest is the model trained by YOLO v5s, whose size is only 52.9 MB, and its mAP@0.5 on the test set and validation set are 0.702 and 0.705, respectively, and mAP@0.5:0.95 in the test and validation are 0.473 and 0.475, respectively. In the case where the size of the model differs by more than ten times, the maximum difference of the mAP does not exceed 8%. From this, we conclude that YOLO v5 can train the corresponding model according to the needs of the researcher. If the accuracy needs to be better and the model size is required, we can choose YOLO v5s for training. If there is no requirement for the size of the model but a requirement for accuracy, we can choose YOLO v5x for training. And in terms of training speed, YOLO v5s is much faster than YOLO v5x.

In order to better test the generalization and robustness of the OD, we also used the YOLO v3 detector to train the OD. The version we used was modified based on the TensorFlow version. The weight parameters are restored from YOLO v3_COCO.ckpt during training. First, the anchor we used is the default anchor of coco: 10, 13, 16, 30, 33, 23, 30, 61,

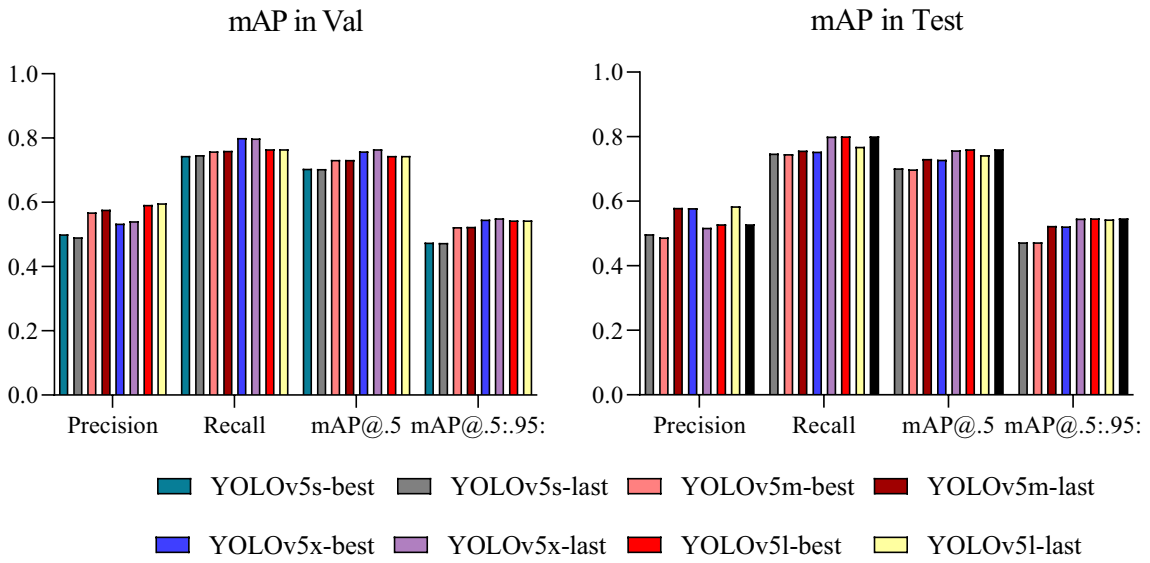

Fig.5 Result of YOLO v5 training the OD

62, 45, 59, 119, 116, 90, 156, 198, 373, 326. The training is divided into two stages: The first stage freezes some layers to train 20 epochs, and the second stage releases all layers to train 43 epochs. Because the size of the object is different from COCO, in order to verify the influence of the anchor of the YOLO v3 on the training difference of the dataset, we used K-means clustering based on the OD to generate a new anchor: 7, 11, 20, 16, 12, 31, 38, 30, 21, 66, 68, 62, 36, 176, 121, 159, 323, 332. Then we use the same method to train to get the comparison model. We initially thought that only 63 epochs were too few, so we continued to train the model up to 100 epochs. When we evaluated it, the AP (average precision) in each category might change, but the mAP did not improve. This also shows that the model at this time has basically reached the optimal state. Figure 6 shows the AP, recall and precision of various categories of the YOLO v3 on the test set and validation set. Based on this training result, we converted the model into .pb file for recognition of videos and images. In Sect. 4.3, we present the test results of our trained model.

In order to contrast with the YOLO-based method, we used the SSD detector training OD. The SSD uses multi-layer features for detection, so the detection effect of small targets is better. We recover the weights from the pre-trained model of vgg16 (Visual Geometry Group Network-16) and use a ssd_300_vgg model to train the OD 100 k steps. Then use a ssd_512_vgg model to fine-tune the trained model of 100 k steps, and total_loss has stabilized and the model is

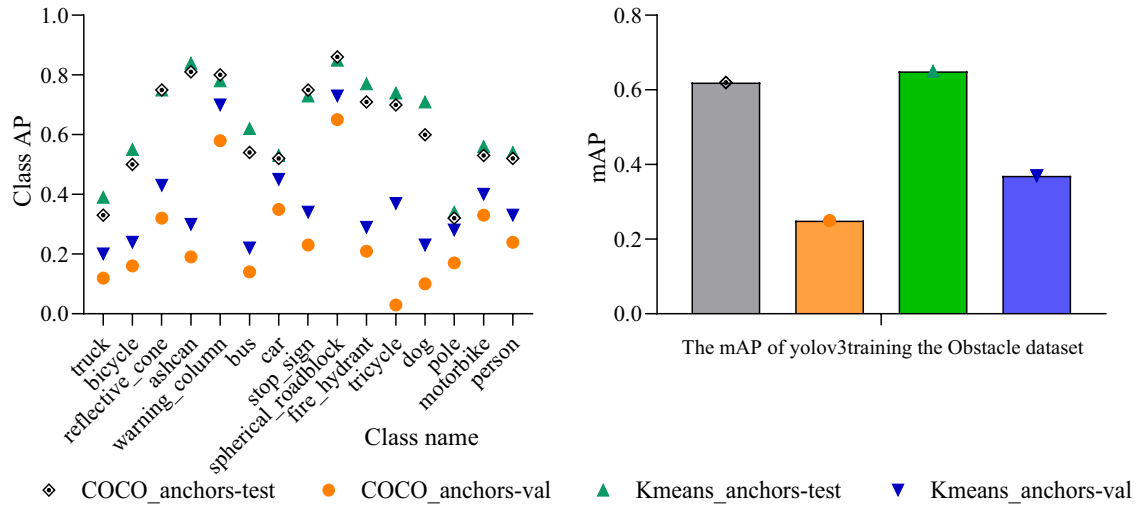

Fig.6 Result of YOLO v3 training the OD

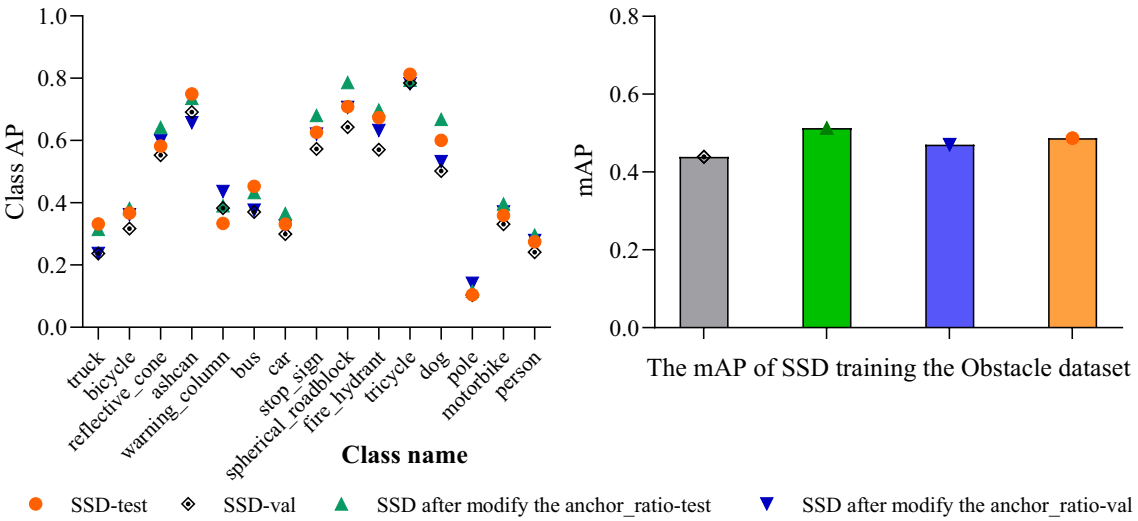

Fig.7 Result of SSD training the OD

basically optimal. Through fine-tuning, we can improve the AP of various categories. However, due to the differences between the anchor set in the SSD and the object in the OD, for the special anchor problem, we modify the anchor_ratio of the ssd_300_vgg and ssd_512_vgg, so that the generated a priori frame can better match the object with a special size to reduce the loss during training. The evaluation results of the SSD-trained model are shown in Fig. 7.

It can be seen from the figure that in the results of training using the original SSD detector, after modifying anchor_ratio, AP (green) in most categories can be improved by 2–3% than before, because SSD sets the ratio of generating a priori frame in the network structure. If the size of the labeled box between our training dataset and the VOC has differences, we can adjust the value of anchor_ratio to match the bounding box with a special size, which will also accelerate the convergence rate of the loss during training to achieve fast training the goal of. We use the two trained models for the actual video and image detection (In Sect. 4.3). Since mAP is lower compared to the YOLO series, the detection effect is also worse, which also shows that the OD established has certain challenges to the SSD algorithm.

The above single-stage detectors are mainstream detectors in recent years, and all object detection needs to be used. In addition to the verification of the above single-stage detector, we also use the two-stage detector Faster RCNN to train the dataset. We use vgg16 and res50 as backbones to train the OD 50 k steps. As shown in Fig. 8, from the evaluation results, it can be seen that the mAP of the model trained with vgg16 is much lower than res50, and from the actual picture detection effect of the picture, the res50 (residual network) trained model is better than the vgg16. However, the evaluation results of the two models are lower than those of SSD and YOLO, which also shows that the OD is also very challenging for some deep convolutional networks.

5.3 Actual detection

In order to verify the actual detection effect, we use the trained model to detect and recognize the image, as shown in Fig. 9.

The selected images contain a total of nine targets. Figure 9a–d shows that all four models trained by YOLO v5 are able to recognize objects of different sizes from small to large in the figure. The detection results of YOLO v3 in Fig. 9e, f show that the difference between the two is not very large. The original SSD can only recognize the closer objects. From the detection results of Faster RCNN in Fig. 9g, h, we can see that the detection effect of using res50 as backbone is better than vgg16. From the detection results of SSD in Fig. 9i, j, we can see that the actual detection effect of the two models is not much different.

Table 2 shows the results of the specific analysis, which includes person × 5, motorbike × 2, tricycle × 1 and pole × 1 in the visual range. It can be seen that the model of YOLO series is strong in detecting the obstacle, which is a motorbike ridden by a person, and the target is missed because its features are not obvious due to its partial occlusion. SSD and Faster RCNN have more missed detection and duplicate detection, which is due to some problems in the network structure design of SSD and Faster RCNN, and the feature extraction ability is not strong. For the detection speed, we use the trained model to detect the same 37 pictures and

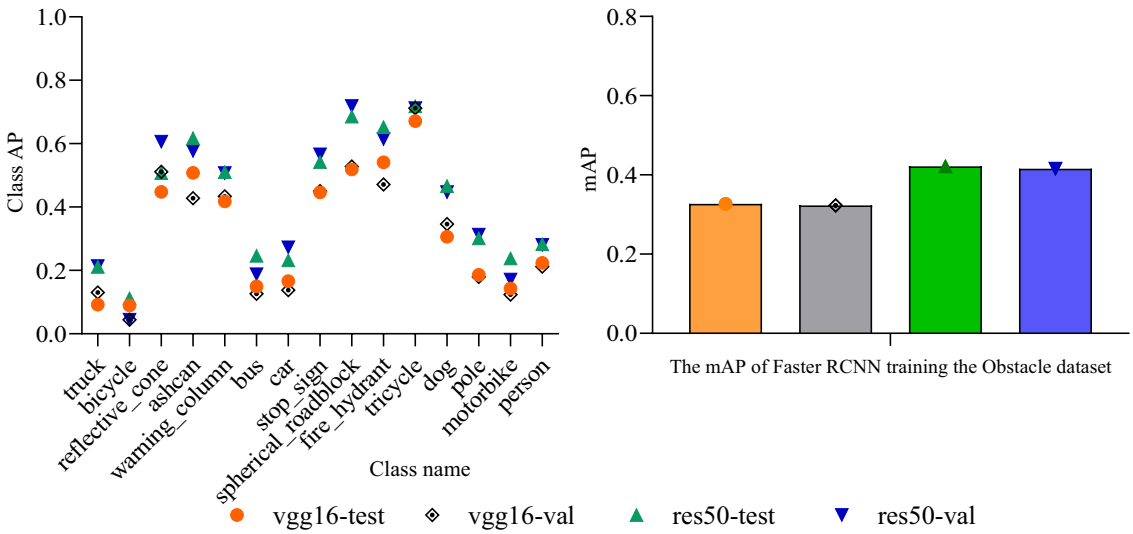

Fig.8 Result of Faster RCNN training the OD

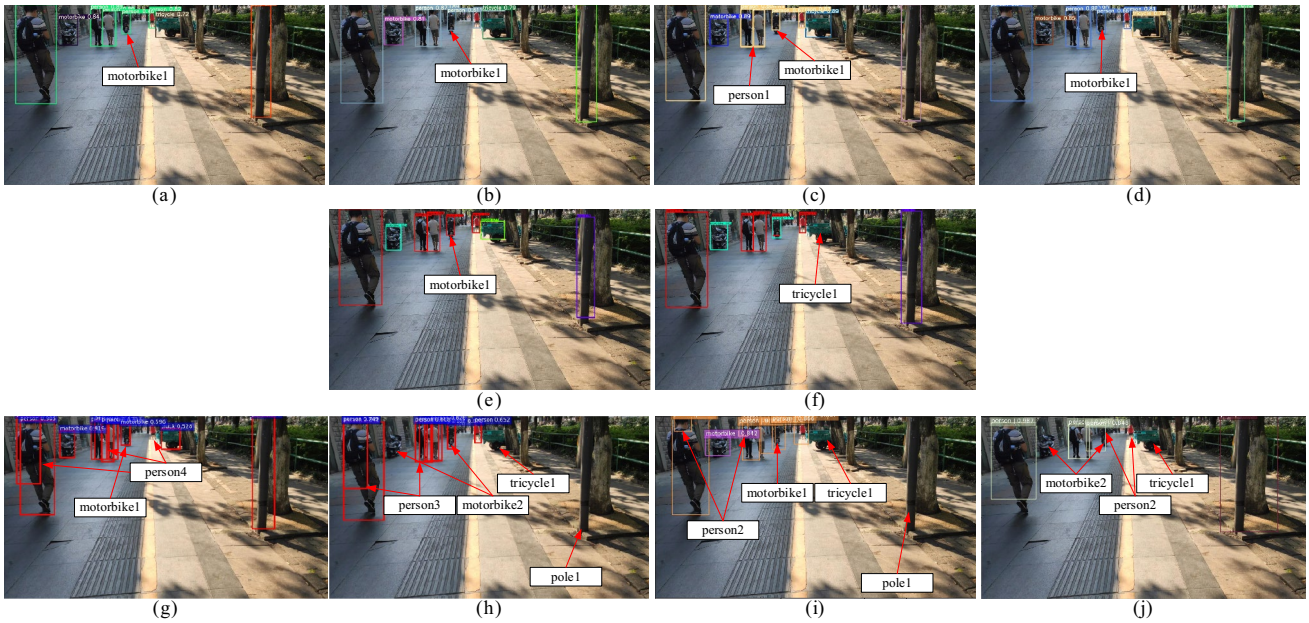

Fig.9 detection results of our trained model on the picture: **a** YOLO v5_s, **b** YOLO v5_m, **c** YOLO v5_x, **d** YOLO v5_l, **e** YOLO v3_coco, **f** YOLO v3_Kmeans, **g** Faster RCNN _res_50, **h** Faster RCNN_vgg_16, **i** SSD_demo, **j** SSD_anchor_ratio_demo, YOLO v5, YOLO v3 and Faster RCNN detection threshold is set to 0.5, SSD threshold is set to 0.9

Table 2 Analysis of detection results of each model

Include		Detection model									
Target	Number	YOLO v5s	YOLO v5m	YOLO v5x	YOLO v5l	YOLOv3-COCO	YOLOv3-K-means	Faster RCNN-res	Faster RCNN-vgg	SSD	SSD _anchor_ ratio
Person	5	5	5	6	5	5	5	7	8	7	3
Motorbike	2	1	1	1	1	1	2	1	0	1	0
Tricycle	1	1	1	1	1	1	0	1	0	0	0
Pole	1	1	1	1	1	1	1	1	0	0	1
Missing detection	–	1	1	1	1	1	1	2	4	3	5
Repeat detection	–	0	0	1	0	0	0	3	3	2	0
Wrong detection	–	0	0	0	0	0	0	0	0	0	0

Table 3 Detection speed

Model name	YOLO v5l	YOLO v5x	YOLO v5m	YOLO v5s	YOLO v3spp	SSD	Faster RCNN	YOLO v3
Speed (s/pic)	0.078	0.149	0.044	0.023	0.098	0.09	0.12	0.10

Table 4 Cross-validation results

Dataset	VOC2007	OD	VOC2007 + 12
SSD	0.421	0.487	–
YOLO v3	–	0.617	0.658

obtain the detection speed of each model, as shown in Table 3.

Table 3 shows that the model detection speed of YOLO v5s is the fastest, reaching 0.023 s per image. Among the trained models, the detection speed of the YOLO5x model is the slowest, with only one image in 0.14 s. The video and image recognition results of all models are available on Google Drive (https://drive.google.com/drive/folders/

Table 5 Comparison with existing works

References	Year	Method	Environment	Class-num	Classes	Established dataset
[47]	2018	Object detection	Outdoor	11	Bench, bicycle, car, dog, motorbikes, person, pole, stair, traffic signals, trees, walls	No
[48]	2018	Object detection	Indoor	2	Chair, table	Yes
[49]	2019	Object detection	Outdoor	7	Bicycle, bus, car, motorcycle, person, truck and traffic light	Yes
[50]	2019	Object detection	Outdoor	2	Person, dog	No
[51]	2019	Object detection	Outdoor	3	Person, cup ball	Yes
[18]	2020	Image classification	Outdoor	3	Other, vehicle, person	Yes
[17]	2020	Object detection	Indoor	6	Person, pothole, car, stairs, chair, washrooms	No
[20]	2020	Object detection	Indoor	16	Window, notice table, elevator, door, electricity box, sign, light, trash can, stairs, security button, table, smoke detector, heating, fire extinguisher, light switch, chair	Yes
[52]	2020	Object detection	Outdoor	12	Background, bicycle, motorbike, bus, car, chair, dog, person, bottle, horse, train, TV monitor	No
[53]	2020	Object detection	Outdoor	90	COCO dataset	No
Our	2021	Object detection	Outdoor	15	Person, bicycle, car, motorbike, bus, reflective, cone, truck, warning column, dog, spherical roadblock, fire hydrant, stop sign, ashcan, tricycle, pole	Yes

1WYJwbKOUQCDoNlSPU6NHaYp5N3FAYhYS?usp=sharing).

6 Cross-validation

Cross-validation of the OD is an evaluation of the generalization ability of the OD. Therefore, we chose the VOC2007 because the VOC2007 contains 20 categories, and the number of categories is close to OD. We selected 5011 images for training and 3221 images for testing. Secondly, we also selected VOC2012 and VOC2007 datasets, using 17,125 pictures in the VOC2012 dataset as a training set and 5011 pictures in VOC2007 as a test set. We used the classic algorithm and YOLO v3 as the object detector for cross-validation. Use SSD to train VOC2007 and OD, and use YOLO v3 to train VOC2007 + 2012 dataset and OD. The labeling method and data format of the dataset are consistent, and the evaluation results are shown in Table 4. The mAP of SSD training models in VOC2007 and OD is 0.421 and 0.487, respectively. Since VOC2007 has more categories than OD, mAP is lower than the OD model. The model mAP trained by YOLO v3 in OD and VOC2007 + 12 is 0.617 and 0.658, respectively. The VOC2007 + 12 dataset contains more images and annotation frames than OD, so its mAP is higher than that of the OD model. The mAP difference between the models trained by the two detectors does not exceed 10%. Therefore, we can think that the OD has good generalization performance as the VOC dataset.

We also list results similar to our study in recent years for comparison with ours, as listed in Table 5. All of these works address the detection of objects for blind people traveling indoors or outdoors. One also addresses image classification, with some proposing their own datasets and some using public datasets. The comparison shows that many methods use a smaller class of datasets and only propose a new detection method and perform experimental analysis on a small-scale dataset they build. We created the largest number of dataset categories in all outdoor datasets. The proposed dataset fills the gap in the blind outdoor travel barrier detection dataset, and we also evaluate and test the dataset using the state-of-the-art target detection method to evaluate the reliability of the dataset. We also make the dataset publicly available for free use by most researchers. This will be a new benchmark for a new method of assessing outdoor travel impairment detection for the visually impaired.

7 Conclusion

We systematically identified some obstacles that are common around blind sidewalks and classified them according to their state attributes. Secondly, according to the classification results, we selected some specific objects to create the obstacle detection dataset OD, which is used to assist the obstacle detection of visually impaired people when traveling. The OD contains 15 common obstacle objects, which will be a new benchmark for outdoor obstacle detection. We used the four object detectors of YOLO, SSD and Faster RCNN to train the dataset to obtain the detection

model. We found that the evaluation results of YOLO v5 performed best overall, and were able to accurately identify and mark the location of objects. The mAP on the test set and validation reached 0.752 and 0.756, respectively, which is 10% and 38% higher than YOLO v3, 24% and 29% higher than SSD, and compared with Faster RCNN, increased by 33% and 34%. Moreover, we applied all models to the actual detection of images and videos, to judge the quality of the model based on the detection effect. At the same time, we used different stages of detector training data, including single-stage detectors and two-stage. Experiments prove that the OD is very challenging for some deep learning models. We also used the VOC dataset and the OD for cross-validation. The difference between the two datasets trained with SSD is only 6%, and the difference between with YOLO is only 4%. The results of cross-validation prove that our dataset has good generalization performance compared to the VOC dataset. Finally, our results are compared with existing research results, and the analysis concludes that our proposed dataset will become a new benchmark for evaluating methods to detect mobility impairments for the visually impaired. We trained and tested with multiple detectors to verify the reliability of the dataset, and our research has its own contribution compared to other existing results, and we open-source the dataset so that more researchers can use the dataset and detection model for free.

In the past few years, the development of computer vision has brought many opportunities and challenges, opened up many scholars to this field of research, and we also need to continue to conduct in-depth research in our field. Blind people have high requirements for obstacle avoidance when traveling outdoors, and it is also important to prompt them of the location of the obstacle after it has been detected and recognized. In the future, we need to focus our research on 3D morphological detection and orientation detection of obstacles, and make full use of this dataset and detection model to minimize the burden of traveling for the blind. At present, we are still paying too little attention to the lives of visually impaired people. It is our hope that more researchers will join efforts to address the needs of persons with disabilities and make good use of resources and technology so that they can enjoy the beauty of the world like everyone else.

Acknowledgements We would like to acknowledge the anonymous reviewers and authors of cited papers for their detailed comments, without which this work would not have been possible. This work was supported by the National Natural Science Foundation of China (Nos. 41361077, 41561085) and the National Natural Science Foundation of Jiangxi Provence, China (No. 20202BAB202025).

References

1. Katika, B.R., Karthik, K.: Face anti-spoofing by identity masking using random walk patterns and outlier detection. Pattern Anal. Appl. **23**, 1735–1754 (2020). https://doi.org/10.1007/s10044-020-00875-8
2. Sajjad, M., Nasir, M., Muhammad, K., Khan, S., Jan, Z., Sangaiah, A.K., Elhoseny, M., Baik, S.W.: Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities. Futur. Gener. Comput. Syst. **108**, 995–1007 (2020). https://doi.org/10.1016/j.future.2017.11.013
3. Zhang, J., Wu, X., Hoi, S.C.H., Zhu, J.: Feature agglomeration networks for single stage face detection. Neurocomputing **380**, 180–189 (2020). https://doi.org/10.1016/j.neucom.2019.10.087
4. Chen, X., Wang, T., Zhu, Y., Jin, L., Luo, C.: Adaptive embedding gate for attention-based scene text recognition. Neurocomputing **381**, 261–271 (2020). https://doi.org/10.1016/j.neucom.2019.11.049
5. Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. (2019)
6. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. arXiv. (2019). https://doi.org/10.1609/aaai.v34i07.6812
7. Hao, Y., Xu, Z.J., Liu, Y., Wang, J., Fan, J.L.: Effective crowd anomaly detection through spatio-temporal texture analysis. Int. J. Autom. Comput. **16**, 27–39 (2019). https://doi.org/10.1007/s11633-018-1141-z
8. Krumm, J.C., Horvitz, E.J., Wolk, J.K.: Localized Anomaly Detection Using Contextual Signals. WO2017048585 A1[P]
9. Song, W., Jia, G., Zhu, H., Jia, D., Gao, L.: Automated pavement crack damage detection using deep multiscale convolutional features. J. Adv. Transp. (2020). https://doi.org/10.1155/2020/6412562
10. Hassaballah, M., Kenk, M.A., El-Henawy, I.M.: Local binary pattern-based on-road vehicle detection in urban traffic scene. Pattern Anal. Appl. **23**, 1505–1521 (2020). https://doi.org/10.1007/s10044-020-00874-9
11. Bu, Q., Yang, G., Ming, X., Zhang, T., Feng, J., Zhang, J.: Deep transfer learning for gesture recognition with WiFi signals. Pers. Ubiquitous Comput. (2020). https://doi.org/10.1007/s00779-019-01360-8
12. Hosni Mahmoud, H.A., Mengash, H.A.: A novel technique for automated concealed face detection in surveillance videos. Pers. Ubiquitous Comput. (2020). https://doi.org/10.1007/s00779-020-01419-x
13. Xiaomeng, C.: A case study on the difficulty of outdoor activities in the college students with visual impairments. J. Suihua Univ. **37**, 1–6 (2017)
14. KR-VISION Technology Co., L.: Krvision, http://www.krvision.cn/offical/page/assist1.html
15. Tapu, R., Mocanu, B., Bursuc, A., Zaharia, T.: A smartphone-based obstacle detection and classification system for assisting visually impaired people. Proc. IEEE Int. Conf. Comput. Vis. 444–451 (2013). https://doi.org/10.1109/ICCVW.2013.65
16. Gorapudi, R., Darsini, P.P., Kavya, U.N., Jaswanthi, O.: Product label, obstacle and sign boards detection for visually impaired people. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3643597
17. Yadav, S., Joshi, R.C., Dutta, M.K., Kiac, M., Sikora, P.: Fusion of object recognition and obstacle detection approach for assisting visually challenged person. 2020 43rd Int. Conf. Telecommun. Signal Process. TSP 2020. 537–540 (2020). https://doi.org/10.1109/TSP49548.2020.9163434
18. Jarraya, S.K., Al-Shehri, W.S., Ali, M.S.: Deep multi-layer perceptron-based obstacle classification method from partial visual

information: application to the assistance of visually impaired people. IEEE Access. **8**, 26612–26622 (2020). https://doi.org/10.1109/ACCESS.2020.2970979
19. Afif, M., Ayachi, R., Said, Y., Pissaloux, E., Atri, M.: Recognizing signs and doors for indoor wayfinding for blind and visually impaired persons. 2020 Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2020. 10–13 (2020). https://doi.org/10.1109/ATSIP49331.2020.9231933
20. Afif, M., Ayachi, R., Said, Y., Pissaloux, E., Atri, M.: An evaluation of retinanet on indoor object detection for blind and visually impaired persons assistance navigation. Neural Process. Lett. **51**, 2265–2279 (2020). https://doi.org/10.1007/s11063-020-10197-9
21. Park, H., Lee, J.: Implementation of an obstacle recognition system for the blind. 2nd ieee eurasia conf. IOT, Commun. Eng. 2020, ECICE 2020. 125–128 (2020). https://doi.org/10.1109/ECICE50847.2020.9302019
22. Park, H., Lee, J.: Implementation and evaluation of obstacle recognition system for the blind. 2nd IEEE Eurasia Conf. IOT, Commun. Eng. 2020, ECICE 2020. 125–128 (2020). https://doi.org/10.1109/ECICE50847.2020.9302019
23. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**, 98–136 (2015). https://doi.org/10.1007/s11263-014-0733-5
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 8693 LNCS, 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48
25. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 743–761 (2012). https://doi.org/10.1109/TPAMI.2011.155
26. Zhang, S., Benenson, R., Schiele, B.: CityPersons: a diverse dataset for pedestrian detection. Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-Janua, 4457–4465 (2017). https://doi.org/10.1109/CVPR.2017.474
27. Braun, M., Krebs, S., Flohr, F., Gavrila, D.M.: EuroCity persons: a novel benchmark for person detection in traffic scenes. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 1844–1861 (2019). https://doi.org/10.1109/TPAMI.2019.2897684
28. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. 1–10 (2014)
29. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: COCO-Text: dataset and benchmark for text detection and recognition in natural images. (2016)
30. Behrendt, K., Novak, L., Botros, R.: A deep learning approach to traffic lights: detection, tracking, and classification. Proc. IEEE Int. Conf. Robot. Autom. 1370–1377 (2017). https://doi.org/10.1109/ICRA.2017.7989163
31. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S.: Traffic-sign detection and classification in the wild. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 2110–2118 (2016). https://doi.org/10.1109/CVPR.2016.232
32. Yucel, M.K., Bilge, Y.C., Oguz, O., Ikizler-Cinbis, N., Duygulu, P., Cinbis, R.G.: Wildest faces: face detection and recognition in violent settings. arXiv. (2018)
33. Nada, H., Sindagi, V.A., Zhang, H., Patel, V.M.: Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. 2018 IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. BTAS 2018. 1–10 (2018). https://doi.org/10.1109/BTAS.2018.8698561
34. Lam, D., Kuzma, R., McGee, K., Dooley, S., Laielli, M., Klaric, M., Bulatov, Y., McCord, B.: xView: Objects in context in overhead imagery. arXiv. (2018)
35. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: a large-scale dataset for object detection in aerial images. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 3974–3983 (2018). https://doi.org/10.1109/CVPR.2018.00418
36. Ta, T.L.: LabelImg, https://github.com/tzutalin/labelImg
37. Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomammana, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznanski, J., Fang, J., Yu, L., changyu98, Wang, M., Gupta, N., Akhtar, O., PetrDvoracek, Rai, P.: ultralytics/YOLO v5: v3.1 - Bug Fixes and Performance Improvements (2020). https://doi.org/10.5281/zenodo.4154370
38. Girshick, R.: Fast R-CNN. Proc. IEEE Int. Conf. Comput. Vis. 2015 Inter, 1440–1448 (2015). https://doi.org/10.1109/ICCV.2015.169
39. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 9905 LNCS, 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2
40. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
41. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: DSSD: Deconvolutional single shot detector. arXiv. (2017)
42. Kolekar, A., Dalal, V.: Barcode detection and classification using SSD (single shot multibox detector) deep learning algorithm. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3568499
43. Du, Y., Pan, N., Xu, Z., Deng, F., Shen, Y., Kang, H.: Pavement distress detection and classification based on YOLO network. Int. J. Pavement Eng. (2020). https://doi.org/10.1080/10298436.2020.1714047
44. Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., Wang, R.: DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. Inf. Sci. (Ny) **522**, 241–258 (2020). https://doi.org/10.1016/j.ins.2020.02.067
45. Zhu, X., Chen, C., Zheng, B., Yang, X., Gan, H., Zheng, C., Yang, A., Mao, L., Xue, Y.: Automatic recognition of lactating sow postures by refined two-stream RGB-D faster R-CNN. Biosyst. Eng. **189**, 116–132 (2020). https://doi.org/10.1016/j.biosystemseng.2019.11.013
46. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLO v4: Optimal speed and accuracy of object detection. arXiv. (2020)
47. Parikh, N., Shah, I., Vahora, S.: Android smartphone based visual object recognition for visually impaired using deep learning. Proc. 2018 IEEE Int. Conf. Commun. Signal Process. ICCSP 2018. 420–425 (2018). https://doi.org/10.1109/ICCSP.2018.8524493
48. Ying, J.C., Li, C.Y., Wu, G.W., Li, J.X., Chen, W.J., Yang, D.L.: A deep learning approach to sensory navigation device for blind guidance. In: Proceedings—20th international conference on high performance computing and communications, 16th international conference on smart city and 4th international conference on data science and systems, HPCC/SmartCity/DSS 2018. pp. 1195–1200 (2019)
49. Zhou, Z., Lan, X., Li, S., Zhu, C., Chang, H.: Feature pyramid SSD: outdoor object detection algorithm for blind people. 2019 IEEE 5th Int. Conf. Comput. Commun. ICCC 2019. 650–654 (2019). https://doi.org/10.1109/ICCC47050.2019.9064251
50. Arora, A., Grover, A., Chugh, R., Reka, S.S.: Real time multi object detection for blind using single shot multibox detector. Wirel. Pers. Commun. (2019). https://doi.org/10.1007/s11277-019-06294-1

51. Shah, S., Bandariya, J., Jain, G., Ghevariya, M., Dastoor, S.: CNN based auto-assistance system as a boon for directing visually impaired person. Proc. Int. Conf. Trends Electron. Inf. (2019). https://doi.org/10.1109/ICOEI.2019.8862699
52. Joshi, R., Tripathi, M., Kumar, A., Gaur, M.S.: Object recognition and classification system for visually impaired. In: Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020. pp. 1568–1572 (2020)
53. Abraham, L., Mathew, N.S., George, L., Sajan, S.S.: VISION: wearable speech based feedback system for the visually impaired using computer vision. In: Proceedings of the 4th international conference on trends in electronics and informatics, ICOEI 2020. pp. 972–976 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.