Image recognition and blind-guiding algorithm based on improved YOLOv3

Haoyu Lu1,* and Yan Ma2

1 School of Software, Zhengzhou University, Zhengzhou, China
2 School of International, Zhengzhou University, Zhengzhou, China

*Corresponding author: lhy13223041056@stu.zzu.edu.cn

Abstract. YOLOv3, a target detection algorithm based on deep learning, is widely applied in object recognition, especially in guiding the blind. The existing products of assisting the blind based on YOLOv3 can already achieve high-precision, high-real-time object recognition. But YOLOv3 also has many limitations, such as the inability to measure distances or it's hard to recognize objects correctly in fog or haze. For these deficiencies, this paper proposes a road barrier monitoring method based on improved YOLOv3, using an image downsampling algorithm based on the dark channel to defog the image, and then with the binocular distance measurement algorithm to calculate the obstacles from the distance of the camera according to the width and height of the obstacles. The experimental results show that the improved product retains the advantages of high accuracy and fast recognition speed of YOLOv3. At the same time, it also owns the new functions of obstacle ranging and bad weather identification. The improved algorithm can meet the requirements of portability, real-time, and practicality of guide products.

Keywords: YOLOv3, Image Recognition, Blind Guide Algorithm.

1. Introduction

Globally, of the 7.33 billion people alive in 2015, an estimated 36.0 million were blinds who are in great demand for guide products [1]. However, there is a lack of assisted travel tools specifically designed for the visually impaired. Take the walking sticks and guide dogs that are most commonly used by people with visual impairments as examples. In real life, most people with visual disabilities can only use ordinary blind canes without technology to assist in walking, which has too many limitations. Although guide dogs can better assist the visually impaired to walk, they also have the disadvantages of long training periods and high costs. In order to solve the above pain points, people have tried many different technologies, and at present, the most potential technology direction of guide products is the blind guidance technology based on image recognition. In this paper, we proposed an improved and comprehensive image recognition algorithm in guiding blindness which combined with the YOLOv3 image recognition algorithm, fog removal algorithm, traffic light recognition, the dual camera ranging algorithm, aiming to help blind people travel more conveniently and safely.
2. Introduction to key technologies

2.1. How to determine the location with YOLOv3
The easiest and most direct method to locate the objection is to use different aspect ratios and sizes to traverse the entire image, but that is too inefficient. Before YOLOv3, the most prevailing way to recognize images was RCNN [2]. It has two steps: first, identify the candidate areas, and then identify the objects in them. It means to search the image for the location of some possible objects, about 2,000. Then object recognition is performed for each candidate area.

And YOLO creatively combines the two stages into one. Yolo doesn't discard candidates but uses predefined candidates. For example, YOLOv1 divides the image into 7*7 grids, YOLOv2 [3] and YOLOv3 [4] divide the image into 13*13 grids. Each grid sets several anchor boxes (equivalent to the candidate areas of RCNN) to determine whether objects exist in the grid. Then, k-means clustering is used to get the optimal size of the anchor frame, and finally, border regression is used to adjust the position of the anchor frame to better fit the target object.

2.2. How to recognize images with YOLOv3
To understand how the machine identifies the image, the first thing to understand is how our eyes work. First, the original signal token by machine (pupils intake pixel), then do preliminary processing (some cerebral cortex cells found the edge and direction), then go abstract (the brain determines the shape of the object), and later go further abstract (the brain further determines what the object is).

So with that out of the way, how do computers recognize images? It's easy for computers to "see" encoded lower layer information such as pixels, but hard to understand higher layer semantic concepts such as whether a target appears in an image or video frame is people or an object, let alone locate the target. The target recognition is applied to make the computer automatically recognize the target category of a picture or video frame and draw the boundary box around the target. The process of image recognition is to classify the original optical information logically.

The steps of image recognition are as follows:

![Figure 1. The steps of image recognition.](image)

The image preprocessing mainly uses the image transformation technology to eliminate the irrelevant information in the image and enhance the useful information to improve the reliability of feature extraction, image segmentation, matching, and recognition. General preprocessing usually concludes image graying, image filtering denoising, image segmentation, image enhancement, image compression, and restoration. Feature extraction and matching are to extract the deep features of the image through a convolution algorithm, including texture feature, color feature, shape feature, and spatial relation. For image training and classification, image classification samples are used to train the convolutional layer, and an appropriate convolution kernel is summarized to extract image features. The combination of a supervised algorithm and an unsupervised algorithm enables the computer to learn the deep features of the target object and improve the recognition accuracy of the algorithm through repeated iteration [5].

2.3. How to speed up with YOLOv3
YOLOv3 performs better than YOLOv2 but slower. The reason is that YOLOv2 uses a custom deep architecture Darknet-19. It was originally a 19-layer network, and there are another 11 layers for object detection. There are either remaining blocks, skipped joins, or up-sampling. YOLOv3 uses a variant of
Darknet, initially training a 53-layer network on Imagenet. To accomplish the detection tasks, 53 layers are stacked on it, providing YOLOv3 with an underlying architecture of 106 layers of complete convolution. It includes remaining blocks, skips joins, and upsampling, which is suitable for recognizing small objects. That’s why YOLOv3 is slower compared to YOLOv2. However, the significant improvement in the recognition effect is well worth the time waste. If it works with a high-performance AI computer such as the Jetson Nano, it can almost offset the defect of time.

2.4. How to improve accuracy with YOLOv3

Compared with YOLOv2, YOLOv3 does not make significant structural changes but uses many practical small methods to improve the operation speed.

Firstly, take different dimensions of the prior box: While the quantity and scale of the output feature map varying, the size of the prior box needs to be adjusted. YOLO2 has used k-means clustering to get the size of the prior box. YOLO3 remains this method, setting three prior frames for each subsampling scale, and clustering a total of nine prior frames. In terms of allocation, large prior boxes are applied on the minimum feature graph (with the largest receptive field), which are suitable for detecting larger objects. The medium-scale box should be applied to the medium characteristic graph (medium receptive field), which is suitable for detecting medium-sized objects. The small prior frame is suitable for a larger feature map (smaller receptive field) and smaller object detection.

Secondly, selects characteristic graphs of different scales during the convolution process: YOLOv3 takes three feature imagines of different scales to detect objects. Compared with the above figure, the convolutional network reaches a scale detection result after 79 layers, passing through several yellow convolutional layers below. Compared to the image input, the feature map used for detection here has 32 times the number of downsamplings. For example, the map's size of the input is 416 x 416, the feature map is going to be 13 x 13. Due to the high lower sampling multiple, the sensitivity field of the feature map is relatively larger, so it is suitable for detecting large objects in the image. To detect the Fine-Grained image, the feature map of the 79th layer started to do upsampling again (starting from the right of the 79th layer, the upper sampling convolution) and then concatenated with the feature map of the 61st layer, to obtain a fine-grained feature map of the 91st layer. After passing through several convolutional layers, we can get the feature map of the input image that was 16 times lower. It has a medium scale receptive field and is suitable for detecting medium scale objects. Finally, the 91st layer feature map upsampling again and concatenated with the 36th layer feature map to obtain a feature map that is subsampled 8 times lower than the input image. It has the smallest receptive field and is suitable for detecting objects of small size.

![YOLO v3 network Architecture](image)

**Figure 2.** The convolution process.
3. Binocular Ranging Algorithm

3.1. Principle of Binocular Ranging

Binocular distance measurement uses the principle of bionics. Two cameras are used to imitate the left and right eyes of a person. The two cameras take two images on the left and right. According to the principle of triangle similarity, we can accurately and quickly extract the distance information of objects from the actual scene. The camera imaging model is shown in the figure below:

![Figure 3. The camera imaging model.](image)

In the experiment, two cameras are placed in parallel. The focal length of the two cameras is \( f \), and the left and right optical centers are represented by \( OL \) and \( OR \). The distance between the left and right optical centers is represented by the letter \( B \). A point \( P \) in the physical space is respectively \( PL \) and \( PR \) on the left and right cameras at the same time. According to the triangle similarity principle, the following formula can be obtained:

\[
\frac{B - (X^L - X^R)}{D - f} = \frac{B}{D} \Rightarrow D = \frac{f \cdot B}{X^L - X^R}
\]

3.2. Camera calibration

The principle of binocular recognition is relatively simple, and the difficulty lies in camera calibration. A special reference object (chessboard paper) placed in front of the camera is usually needed for calibration. The camera can obtain the image of the object and calculate the internal and external parameters of the camera. The position of each feature point on the calibration reference object relative to the world coordinate system should be accurately measured when making, and the world coordinate system can be selected as the object coordinate system of the reference object. After getting the projection positions of these known points on the image, the internal and external parameters of the camera can be calculated [6].

4. Defogging Algorithm

4.1. The purpose and significance of image defogging

In real life, heavy fog or haze weather often occurs in autumn and winter, which not only obstructs people's vision but also affects travel. This is because the intensity and wavelength of the light will change after the light is refracted and scattered under heavy fog or haze, which will seriously lead to the coverage of key information of the collected image. At present, the main image defogging algorithms are mainly divided into two categories. They are methods based on physical defogging models and methods based on image enhancement. The representative algorithm of the former is the defogging
algorithm based on the dark channel proposed by Dr. He Kaiming. The latter's representative algorithms include a defogging algorithm based on multi-scale Retinex image enhancement and an image defogging algorithm based on adaptive contrast and level enhancement.

4.2. Dark channel prior theory
In nature, shadow and dark color are everywhere, which is the main reason for the low channel value of the dark primary color of images. Therefore, it is considered that there will always be some very low values in some color channels in any image region in nature. But the color of fog is always gray-white, which makes the original dark image become gray-white under the influence of fog or haze. This is the essential law that can effectively remove fog. According to the physical model [8] of fog graphics, currently, we only know the image I(x) we have collected, we must find a way to find the atmospheric light component A and transmittance t(x). After experimental and theoretical verification, we finally got the following formula.

$$J(x) = \frac{I(x) - A}{\max(t(x), t_0)} + A$$

Through the above formula, we can process the foggy image, and finally get the fog-free image, which is convenient for subsequent processing.

4.3. Summary of defogging algorithm
The algorithm in the traditional blind guidance system does not take into account the adverse effects of bad weather on the algorithm analysis. So we made the following improvements: before officially identifying the zebra crossings and traffic lights, defogging was done. Considering the real-time needs of the blind guide system, we used the image downsampling defogging algorithm based on the dark channel. In this algorithm, the size of the window and the value of $\omega$ are the main factors affecting the defogging effect. According to the empirical value of the experiment, we decide to take the size of the window as 15, and for the effect of the depth of field of the image, the algorithm takes the value of $\omega$ as 0.9. Finally, the defogging effect and real-time requirements of the defogging module of the blind guidance system are achieved.

5. Conclusions
According to the statistics of the World Health Organization, as of 2015, there were 36 million people with vision impairment in the world. But today, travel is still a great challenge for these blind people. The algorithm proposed in this paper is based on the current popular yolov3, combined with binocular ranging and defogging algorithm based on dark channel theory. It not only has fast recognition speed, less resource consumption, high real-time performance but also is suitable for heavy fog, haze, and other bad weather. We hope that the algorithm can be better applied to blind guidance product development. We hope to provide care for this vulnerable group, provide them with safe, reliable, and low-cost services, and bring light and warmth to 36 million blind people around the world.

References
[1] Bourne R R A, Flaxman S R, Braithwaite T, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis [J]. The Lancet Global Health, 2017, 5 (9): e888-e897.
[2] Girshick R. Fast R-CNN [J]. Computer ence, 2015.
[3] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger [J]. 2017:6517-6525.
[4] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement [J]. arXiv e-prints, 2018.
[5] YU DeJian. Research on Key Algorithms of Guided Blind System Based on Machine Vision [D]. (In Chinese).
[6] Hu Han-ping. Research on Key Technology of Binocular Stereovision Ranging [J]. Knowledge output of Changchun Institute of Optics and mechanics, Chinese Academy of Sciences, 2014. (In Chinese).

[7] He K, Sun J, Fellow, et al. Single Image Haze Removal Using Dark Channel Prior [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33 (12): 2341-2353.