Evaluation of People Detection with SSD from Fisheye Images

Takeru Tanaka^{a,*}, Hideaki Katayama^a

^aNational Institute of Technology, Maizuru College, Maizuru 625-8511, Japan

*Corresponding Author: a0381@g.maizuru-ct.ac.jp

Abstract

In this research, in order to solve the problems visually impaired persons have when they go out, we have developed a system which gives information on three-dimensional space to them by using image processing. As a method for efficiently inputting a three-dimensional space into the system, we took up the acquisition of images with a fisheye lens. So we conducted experiments to examine the usefulness of fisheye image input for general object detection algorithm. And we also performed experiments on the shortening of processing time with a view to mounting it on small devices. From the results, it was found possible, by correcting the distortion aberration, to use the fisheye image for the inputting of the algorithm of object detection.

Keywords: Visually impaired, Camera calibration, Single Shot MultiBox Detector.

1. Introduction

First, we investigated the needs of visually impaired people in daily life. From this survey, we have found the problem to be solved and propose a system that satisfies their needs.

2. Investigation result

2.1 Documentation and Inquiry

The total number of persons with disabilities is 860.2 million, of which the physically handicapped account for the highest percentage of 3,937 thousand people. The total number of persons with disabilities is 860.2 million, of which the physically handicapped account for the highest percentage of 3,937 thousand people. There are 315 thousand visually impaired persons, amounting to about 8%

of the disabled people¹⁾.

In the survey on “Types of obstacles for those who are out or intend to go out” carried out for each type of disability in 2006, the visually impaired answered by choosing the following two from multiple items: “it is inconvenient for me to use public transportation”, and “I feel uneasy about a crowd and cars”. The two choices showed the highest percentage of 32 respectively among visually impaired persons. Moreover, the ratio of visually impaired people who chose the two items is higher than that of people with other disabilities who did so. It can be said, therefore, that grounds for concern specific to the visually impaired are included in the two items. “Lighthouse”, a general welfare facility for the visually handicapped, recommends that non-handicapped people guide a visually disabled person to a seat and to confirm the destination in a bus or a train.

For further investigation, we asked for cooperation from Mr. Kozo Nonomura, a staff member of Kyoto Lighthouse, and Mr. Yoshihiko Umeki who established the association of Kyoto Deaf Blog Hohoemi. As a result of interviews by telephone and e-mail, issues concerning public transportation were brought to the fore, particularly anxiety felt by disabled persons. The main reason is that it is difficult for them to grasp the in-car environment, that buses and trains are not fully equipped with support system including voice guidance, and that few passengers know exactly how to treat them. Investigations have brought to light the fact that they need support when using public transportation.

2.2 Government's efforts

The budget for measures for disabled persons in FY2008 was 1,814,427 million yen, of which “livelihood support” and “insurance & medical care” account for 87.3%³⁾. The budget for public transportation, which is included in that

of “living environment”, is 98 million yen. Specifically, as part of the formulation of guidelines concerning passenger facilities, vehicles, etc., and as part of movement support, the government is promoting pedestrian movement support using ICT. In addition, the Ministry of Land, Infrastructure and Transport has created an open data site on support service for pedestrian movement, and it discloses barrier-free data of facilities for passengers at railway stations and those of buildings used by an unspecified number of people⁴⁾.

2.3 Survey Summary

According to our survey so far, no research on the use of bus by the visually impaired was reported. The government has been working on the spread of non-step buses. The goal set in 2011 by the Cabinet Office concerning barrier-free buses is said to be achieved in 2020. Long-term efforts and a huge cost are often required to install facilities and to cultivate social manners. By dynamically utilizing support services rather than waiting for the improvement of public facilities, we can respond to individual needs at a lower price and in a short period of time. Based on the above survey, we have considered a vacant seat detection system used by a visually disabled person in the bus.

3. System overview

The system recognizes the state of passengers and seats from the images captured in the in-car environment and provides information to users. For object recognition, object detection algorithm using depth learning is used. Fisheye lens is used to obtain the fisheye image used for input this time. Since it has a wider field angle (100 to 180 degrees) than a usual lens (angle of view 25 ° to 50 °), it is possible to accommodate a wide in-vehicle environment in one image. On the other hand, characteristic distortion (hereinafter referred to as distortion aberration) is one of the features. A large distortion occurs toward the outside of the image, and the shape of the object of detection changes, which may make detection difficult. For this reason, the object is detected from the image in which the distortion aberration of the fish-eye lens image is corrected.

4. Theory

4.1 Calibration of image distortion

A pinhole camera and lens distortion are used as the camera model. Fig. 1 shows the pinhole camera model. Camera parameters include internal parameters, external parameters and distortion coefficients. By estimating these parameters, it is possible to make the points of the three-dimensional world coordinates correspond to the points on the two-dimensional image. When moving the world coordinates $M = (X_w, Y_w, Z_w)$ to the image coordinates $m = (u, v)$ with the focal length set to 1, equation (1) is used. $[R | t]$ in Eq. (1) is called an external parameter and represents translation and rotation. This indicates the position of the camera in the three-dimensional space, and converts world coordinates into camera coordinates, which is then projected onto the image plane by an internal parameter A composed of the image center (c_x, c_y) and focal length (f_x, f_y) .

$$m = A[R|t]M \quad (1)$$

Next, since there is no lens in the pinhole camera, distortion coefficient is taken into account in order to express the actual camera. The distortion aberration of the fisheye image can be expressed by the lens distortion in the radial direction and the circumferential direction. In the radial distortion, the refractive index of light increases from the optical axis toward the edge of the image. Circumferential distortion occurs when the lens and the image plane are not parallel. When camera coordinates are (x, y) , the pixel position (x', y') considering the distortion is shown by the following equation :

$$\begin{cases} x' = x(1 + k_1r^2 + k_2r^4 + k_3r^6) + 2p_1xy + p_2(r^2 + 2x^2) \\ y' = y(1 + k_1r^2 + k_2r^4 + k_3r^6) + p_1(r^2 + 2y^2) + 2p_2xy \end{cases} \quad (2)$$

Fig. 1. Pin-hole camera model

where $r^2 = x^2 + y^2$. Using these relationships, distortion aberration is corrected. This time we have adopted Zhang's method which is famous as a calibration method. In Zhang's method, firstly, a checkerboard is photographed with a fisheye lens. Then, by detecting the intersection point, the degree of distortion between orthogonal straight lines is examined. Thus it is possible to estimate the distortion correcting parameter. In this research, Zhang's approach was adopted from the viewpoint that it is not necessary to obtain strict correction and that it can be done easily with a checkerboard.

4.2 Object detection algorithm

An image in which the distortion aberration is corrected by the method shown in the preceding paragraph is used as a corrected image. By using this corrected image, persons are detected in order to grasp the number of passengers. Since detection precision and detection speed are required, a Single Shot type detector is used. While YOLO (You Only Look Once) and SSD (Single Shot Multi Box Detector) are typical examples of this system, we have adopted SSD because it is the quicker and more accurate of the two.

Here we describe the features of object detection of SSD. One of the features of SSD is that it has the rectangular "frame" called "default box". When inputting an image into SSD and predicting what is in it and where it is, SSD places 8732 default boxes of different sizes and shapes on the image and calculates predicted values for each frame. The functions of this default box are as follows. First, it is to evaluate how far from the object the default box is and how big the object is. Then it predicts what is in the default box. There functions are called location prediction and class prediction respectively. In location prediction, the difference based on the coordinates of the corresponding default box is evaluated. In class prediction, probability is expressed as to which object is likely to be surrounded by the frame.

Now, let us explain the selection of the default box. The following operation is carried out on the above two evaluation criteria. First, select the frame with the highest class prediction value. Secondly, if the area overlapping with other frames is not less than 50% of the cumulative area of the two frames, these frames are excluded. This indication of the degree of overlap is called Intersection over Union (IoU) while the above operation is called Non-maximum suppression. The prediction of position and

class above and the selection of the default box constitute the theory of SSD object detection.

5. Experimental environment

Table 1 shows the experimental environment. We have attached a fisheye lens (Fuleadture B 075 WRPNGS) to iphone 6 to take photographs.

6. Experiment

6.1 People detection from corrected image

In order to evaluate the accuracy of fisheye image input to SSD, people detection was conducted with 136 fisheye images (3264×2448 [pixel]) in which one person appears and their corrected images. An example is shown in Fig. 1. We regarded the case in which the person was detected as success. The rest were regarded as failure. The detection rate was 73.5% before correction and 94.1% after correction. Although the detection accuracy increased drastically by correcting the distortion aberration, it was found that the correction requires 4 seconds or more processing time.

6.2 Reduced processing time

Therefore, we have tried to shorten the processing time by degrading the resolution. The processing time is the average of the time required per image. In order to investigate how far we can reduce the processing time by degrading the resolution of the input image and how the detection rate changes, we have changed the number of pixels per side up to 0.01 times with respect to the fisheye image. The experimental results are shown in Fig. 2. It can be seen from this figure that the processing time becomes shorter as the resolution decreases. The detection rate of 90% or more can be secured up to 286×214 [pixel].

Table 1. Experimental environment

CPU	Intel® Core™ i7-4790
OS	Windows8.1
memory	8.00GB

Fig. 2. Detection rate and Processing time

7. Consideration

We extracted needs of visually impaired persons from the result of a survey on their current situation and conducted an experiment to realize a system that satisfies them. Since it was possible to detect a person by correcting the distortion aberration of the fisheye image and inputting it to the SSD, it is concluded that fish-eye image input is useful for the system. Also, assuming implementation on a small device, it is expected that the processing time will increase as the PC specification decreases. Therefore, as a result of experiments on the relationship between resolution and detection rate, it was possible to obtain the detection rate of 91.2% even when degraded to the resolution of 286×214 [pixel].

8. Conclusion

In the future, we plan to examine the method of detecting vacancies. In order to perform detection by SSD as we did for people detection, a data set of a seat image will be created and its learning will be attempted. Since the seat itself has various shapes, there is concern that satisfactory accuracy cannot be obtained with only the object detection algorithm. So we would like to consider a method for assisting the detection of vacant seat position from the number of passengers and their positional relationship.

References

- (1) Cabinet Office : “Report on people with disabilities in 2013” [online] http://www8.cao.go.jp/shougai/whitepaper/h25hakusho/zenbun/pdf/h1_01_01_01.pdf
- (2) Ministry of Health, Labour and Welfare Disability Health and Welfare Department : “Survey result of physically handicapped” [online] <http://www.mhlw.go.jp/toukei/saikin/hw/shintai/06/dl/01.pdf>
- (3) Cabinet Office : “Budget on policies for people with disabilities” [online] <http://www8.cao.go.jp/shougai/suishin/yosan/pdf/yosan.pdf>
- (4) Cabinet Office : “Building a foundation for a living environment” [online] http://www8.cao.go.jp/shougai/whitepaper/h26hakusho/zenbun/pdf/s7_1_4_1.pdf