Object Detection and Distance Estimation Tool for Blind People Using Convolutional Methods with Stereovision

[1)]Rais Bastomi, [2)] Firza Putra Ariatama, [3)] Lucke Yuansyah Arif Tryas Putri, [4)] Septian Wahyu Saputra, [5)] Mohammad Rizki Maulana , [6)] Mat Syai'in, [7)] Ii Munadhif, [8)] Agus Khumaidi, [9)] Mohammad Basuki Rahmat, [10)]Annas Singgih Setiyoko, [11)] Budi Herijono, [12)] E.A. Zuliari, [13)] Mardlijah
[(1-8)]*Automation Engineering Study Program*
[(9-10)]*Marine Electrical Engineering Study Program*
[(1-11)]*Shipbuilding Institute of Polytechnic Surabaya*
[(12)]*Adhi Tama Institute of Techhnology Surabaya*
[(13)]*Sepuluh Nopember Institute of Techhnology*
Surabaya, Indonesia 60111
raisbastomi@gmail.com

***Abstract*— in this research, a tool that can provide information about object around is made. This tool can also estimate distance of detected object through camera which is combined with glasses, to ease blind people who use it. This tool is certainly can help them to identify object around and improve their skill and ability. This tool use camera as main sensor, which works like human eyes, to provide real time video as visual data. The RGB visual data is processed using *Convolutional Neural Network* which has 176x132 pixels by convoluting 2 times. It produces smaller pixels with size 41x33 pixels, so weights is obtained for classification using back propagation and determined dataset. After getting detection result, the next step is a find centroid value as center point for measuring the distance between objects and cameras with *Stereo Vision*. The results is converted into sound form and connected to earphones, so blind people can hear the information. The test results show that this tool can detect predetermined objects, namely humans, tables, chairs, cars, bicycles and motorbikes with an average accuracy of 93.33%. For measurements of distances between 50 cm to 300 cm it has an error of around 6.1%.**

Keywords— *Blind, Centroid, Convolution Neural Network.*

I. INTRODUCTION

Nowadays, object recognition is using essential Machine Learning methods. To improve its performance, we able to collect larger data sets, learn stronger models, and use better techniques to prevent over fitting. But in reality, objects show considerable variability, so to recognize them needs to use larger training set [1]. The dataset containing small numbers of images does indeed have drawbacks, but now it is possible to collect large amounts of data. Such datasets consist of hundreds of thousands of images that are fully segmented, and consist of more than 15 million high-resolution images labeled in more than 22,000 categories. To learn about thousand objects from million images, we need a model with a large learning capacity. And CNN has far fewer connections and parameters so they are easier to train, CNN training in the form of the latest data collection containing sufficient data in the process of training objects without severe over fitting [2]. This network contains two convolutional layers and three layers that are fully connected, and this depth seems important. We find that eliminating any convolutional layer produces lower performance.

In this paper, we train an object detection and distance aid for blind people. Because the blind people have difficulties to identify their environment around them. Several older research about object recognition for blind people has been done. One of them is object detection for blind using SIFT to identify and locate object in a video scene using SIFT and convert the output into sound, but the object label must be towards to camera [3]. It is very tough to blind people to locate label efficiently. The second research is also using SIFT, and can identify object in a picture with many object [4]. From those research, object detection result only provide information about kind of object or convert it into sound. With the result of that, we have an innovation to make object detection for blind people, which added stereo vision in the tools. The object detection and classification uses the CNN method with two convolution processes. The object distance also measured using stereo vision. The result will be connected to the earphone in the form of sound.

Our model has several contributions. First, we define object detection as a regression problem for the environment around the user which consists of many objects. Each predicted box has a distance value measured from the user's position. The second main contribution is combining object detection using CNN with stereo vision for object measurement. And the third contribution is converting the results of the classification and distance measuring process of objects into sound. So, user can hear the objects around him.

II. METHODOLOGY

The purpose of this research is to create a device that can detect an object and determine distance using the Convolutional Neural Network method. The camera is used as the main sensor to get image input. The next stage of this research is explained by the flowchart in figure 1.

978-1-7281-2156-7/19/$31.00 ©2019 IEEE

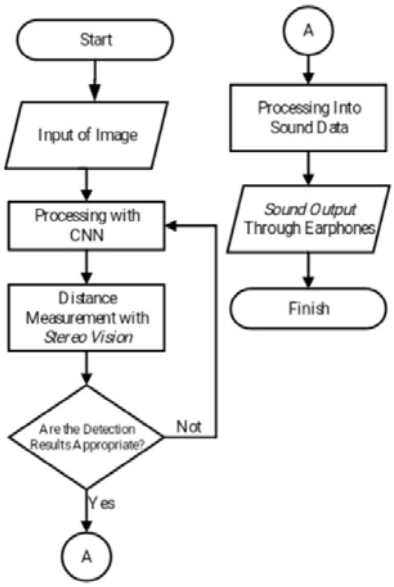

Fig 1. System Flowchart

A. Input Image

Image input is obtained using a webcam camera connected via USB port. It processed using Open Cv, then set into 640 x 320 pixels. Image input is accessed in streaming or real-time, then processed with a convolutional neural network.

B. Convolutional Neural Network

CNN is learning system on a very deep machine. Where is the position equivalent to Deep Learning. CNN is used in image processing, where the weight obtained from the image is classified by the weight of the dataset image. CNN is generally divided into 2 main layers, feature extraction layer and classification layer [4]. The feature extraction layer of this research uses 2 times image convolution to get the image size of 41 x 33 pixels.

1. Feature Extraction Layer

   In this layer there is a convolution layer and max pooling layer. The convolution layer is used to shrink the pixel size of the image. While the max pooling layer is used to maximize the sharpness. So the small image is clearly visible [5]. In this study using 2 times the convolution layer and 2 times the max pooling layer.

A. Convolution layer

This convolution layer uses the Gaussian kernel algorithm. The kernel was chosen because its ability to reduce images by producing less noise in new images [6]. The results of the first convolution is convert image size 176 x 144 to 172 x 140. The second convolution is convert of the first max pooling results with a size of 86 x 70 into 82 x 66.

Convolution data result is a spatial from equation (H). The convolution input data is transformed into linear (I) transformation and has a result [7].

$$I' = H \otimes I$$
$$I'(x,y) = \sum_{i=-n}^{a} \sum_{j=-n}^{b} h(i,j)I(x+i,y+j) \quad (1)$$

The a and b is a measure of the filter function in the matrix. For x and y variables is the size of the x and y values in the pixel image. As in Figure 2 is an example of input and image results from the first convolution process.

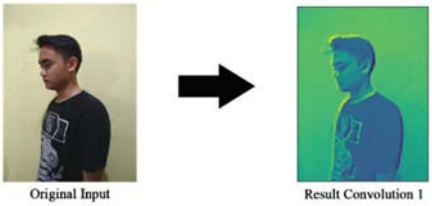

Fig 2. Convolutional image result

B. Max Pooling Layer

Max pooling is used after every image convolution process. First convolution result with a size of 172 x 140 pixels is processed with max pooling to 86 x 70 pixels. These results are used as input for the second convolution. At the second max pooling, the input from the second convolution result is 82 x 66 is processed into 41 x 33. Figure 3 is an example of input from the result of convolution 1 and the image results from the first max pooling process.

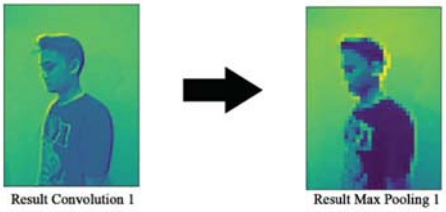

Fig 3. Max Pooling Result

2. Classification Layer

   In this layer, the feature extraction layer result will be classified with the specified dataset value. There are many neurons that are fully connected, one layer or neuron in another layer. Feature extraction layer result is trained using NN (Neural Network) with the Back propagation Neural Network (BPNN) method. The neuron data in NN has added of a sigmoid function to change the range of input values [8].

C. Centroid

After getting bounding-box with the object name as CNN detection result, centroids on the bounding-box is used to measure the distance between objects and cameras. On image processing each form consists of pixels. Centroid is the average arbitrary pixel of the x value, y image. To find the centroid image using conversion to binary format [9].

$$c = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (2)$$

a. n = number of data

X = pixel of x value

D. Stereo Vision

After obtaining the centroid value, a calibration is needed to obtain the variables needed by stereo vision. The most important thing is that the two cameras must be parallel and have the same type. Calibration process use chessboard that

978-1-7281-2156-7/19/$31.00 ©2019 IEEE

has a size of 10 x 8 with 20 x 20 mm grid size as object. The object must be located in the area that can be seen by both cameras by taking the pattern automatically 25 times simultaneously. The results of shooting are as shown in Figure 4 below.

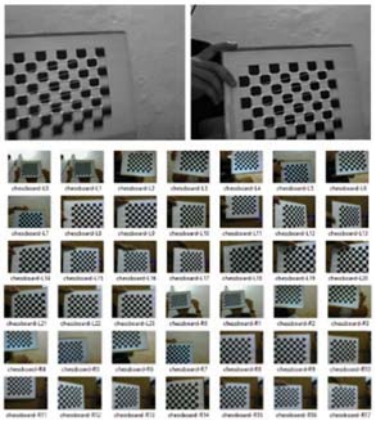

Figure 4. Calibration process

After obtaining the necessary variables, then it used to calculate distance with the stereo vision equation. The following illustration for the stereo vision equation can be seen in Figure 5.

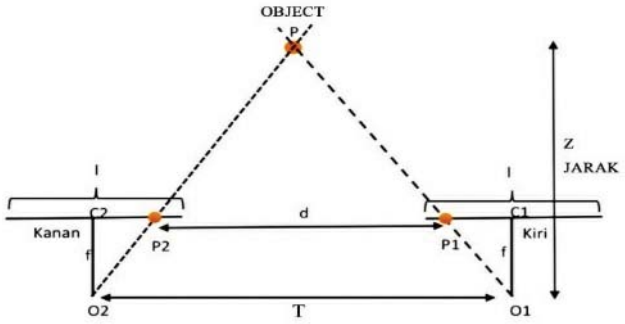

Fig 5. Stereo Vision Concept Illustration

From Figure 5, the stereo vision formula is obtained so that it can estimate the distance of the object [10]. There is the equation used in measuring the distance of objects to the camera with an explanation as follows.

fl = *focal length* of left camera
fr = *focal length* of right camera
T = distance between two camera center points
X_l = centroid point X left camera
X_r = centroid point X right camera

The difference between centroid X frame points like equation 3. And can used to measure the distance as a divider in equation 4 below.

$$d = X_l - X_r \quad (3)$$

$$Z = \frac{f\ x\ T}{d} \quad (4)$$

In this research, the distance detection distance is limited by a minimum distance of 50 CM and a maximum distance of about 300 CM. Above it cannot measure distance properly.

E. Dataset Image

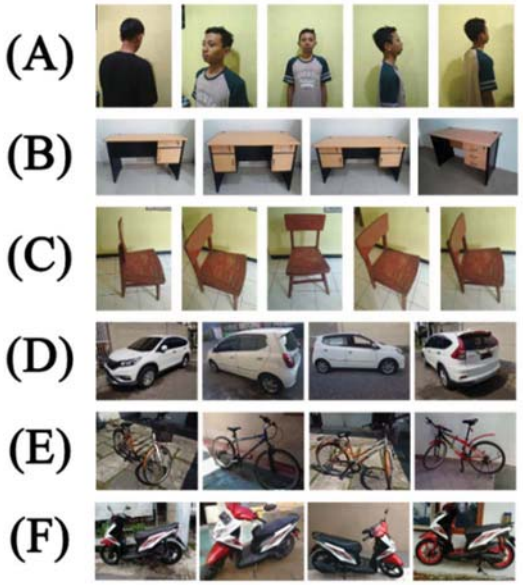

Fig. 6 Dataset Image

Datasets preparation is needed for object recognition systems. There are 6 objects that can be identified by objects as in Figure 4. For image A is a human, B is a table, C is a chair, D is a car, E is a bicycle, and D is a motorcycle. The total content of the dataset was 750 data. Each object has 125 data content. With 100 content as training data and 25 content as testing data. Before entering the training dataset, the training data experiences wrapping and cropping. The examples of images used by the dataset is in Figure 6.

F. Hardware

The main hardware used is the Mini PC, Camera and Battery. This tool is divided into two parts. The first part is a computer for main of the algorithm processing. For the second part there is a camera placement. The camera is combined with the glasses so that the system works like the eye, as in Figure 8.

Fig 7. Glasses with Camera

In Figure 8, the camera is placed on the right and left side of the glasses. There are 2 holes in which there is a camera. The part is designed to resemble glasses so that it can be comfortable.

III. TEST RESULT AND DATA ANALYSIS

In this chapter, is provide real time object detection test result, test results regarding image processing parameter responses, and stereo vision measurement result that compared with a meter.

978-1-7281-2156-7/19/$31.00 ©2019 IEEE

A. CNN Test Result

This test is detecting all objects that have been determined with the dataset in real time. The test result are below.

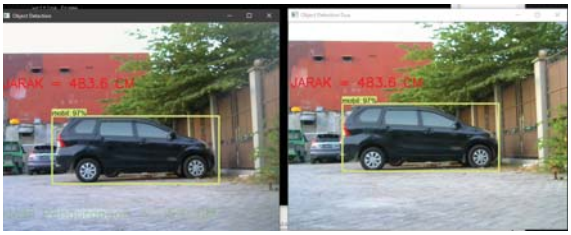

Fig 8. Car Detection Result

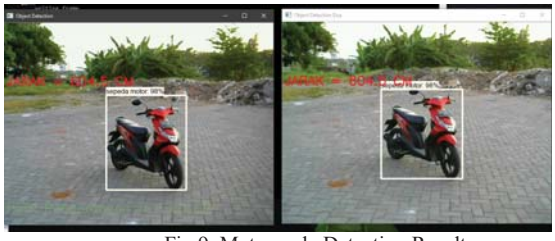

Fig 9. Motorcycle Detection Result

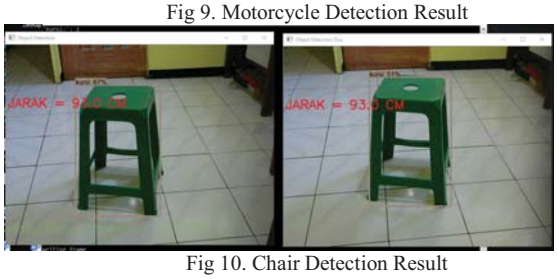

Fig 10. Chair Detection Result

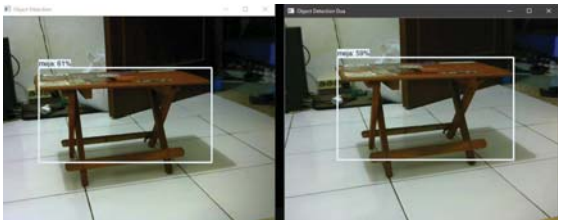

Fig 11. Table Detection Result

Fig 12. Human Detection Result

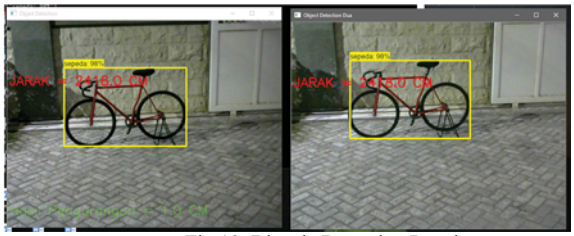

Fig 13. Bicycle Detection Result

In Figure 8, the car can be detected from the side in full. In Figure 9, there is a detection result for a motorcycle. In Figure 10, testing is carried out on chairs made of plastic. Whereas in Figure 11 the table used for testing is small and made of wood. The results of human detection have no obstacles seen in Figure 12. And the last object that can be detected is a bicycle can be seen in Figure 13. Figure 14 shows the configuration results using Hidden Neuron 256. In the graph the accuracy with Hidden Neuron 256 is more stable starting at epoch 10. With this epoch 60 the iteration value remains stable between the ranges of values close to 1.

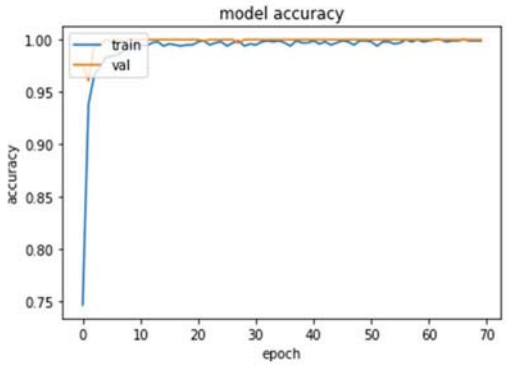

Fig 14. Graphic images regarding performance accuracy

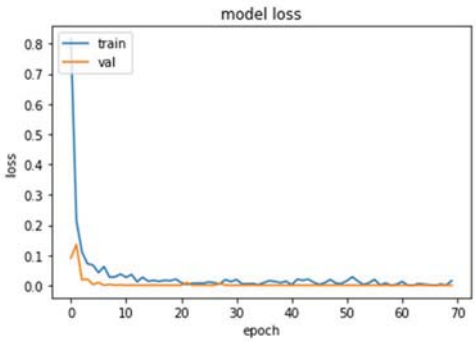

Fig 15. Graphic image of MSE performance

In Figure 15, the average error value starts in the range 0.15 and closer moves to 0.00, after the 20th iteration the average error in each iteration is 0.005 and is stable until the 70th epoch.

B. Stereo Vision Test

Real-time stereo vision measurement use meter comparison technique. Measurements using the camera to the object are carried out on 2 objects, namely human and chair Measurements were made 24 times with 12 times for humans and 12 times for chairs. The results of the test can be seen in Table 1 below.

TABLE 1. HUMAN DISTANCE MEASUREMENT RESULT

No	Meter (cm)	Camera (cm)	*Error* (%)
1	50	52,5	5
2	70	69,2	1,1
3	90	93,2	3,5
4	110	108,3	1,5
5	130	128,6	1
6	150	144,4	3,7
7	170	172,1	1,2
8	190	193,4	1,7

978-1-7281-2156-7/19/$31.00 ©2019 IEEE

9	210	236,4	12,5
10	230	241,2	4,8
11	250	275,3	10,1
12	270	316,2	17,1
Error Percentage			$\underline{x}$ = 5,3%

TABLE 2. CHAIR DISTANCE MEASUREMENT RESULT

No	**Meter (cm)**	**Camera (cm)**	***Error* (%)**
1	50	56,9	13,8
2	70	71,6	2,3
3	90	82,1	8,8
4	110	108,3	1,5
5	130	134,5	3,5
6	150	159,2	6,1
7	170	177,3	4,3
8	190	196,6	3,5
9	210	246,3	17,3
10	230	236,4	2,8
11	250	278,6	11,4
12	270	288,9	7
Error Percentage			$\underline{x}$ = 6,9%

TABLE 3. STEREO VISION ERROR AVERAGE

Object	**Error (%)**
Human	5,3
Chair	6,9
Error Percentage	$\underline{x}$ = 6,1%

In Table 3, the results of testing stereo vision have a considerable error with a value of 6.1%. In testing it is necessary to place a straight object between two cameras in order to measure it properly.

IV. CONCLUSION

2. CNN system performance when detecting objects in the form of cars, tables, chairs, bicycles, humans and motorbikes has their respective characteristics from various directions. There are still unsuitable detection results. According to the results the success rate of detection is quite high at 93.33%.
3. Stereo Vision measurement results have a high error around 6.1%, while the object is located right between 2 cameras. Stereo Vision has a weakness, if the object detected is not right in front of and between the two cameras, the Stereo Vision measurement cannot measure properly. Stereo Vision has limited distance in measurement. Distance measurement is only in the range of 50 cm to about 300cm.

REFERENCES

[1] K. Jansen and H. Zhang, "Scheduling malleable tasks," *Handb. Approx. Algorithms Metaheuristics*, pp. 45-1-45–16, 2007.

[2] P. Hansson, "Fracture Analysis of Adhesive Joints Using The Finite Element Method," *Lund Inst. Technol.*, vol. 60, no. February, pp. 84–90, 2002.

[3] V. Mohane and C. Gode, "Object recognition for blind people using portable camera," *IEEE WCTFTR 2016 - Proc. 2016 World Conf. Futur. Trends Res. Innov. Soc. Welf.*, pp. 3–6, 2016.

[4] H. Jabnoun, F. Benzarti, and H. Amiri, "Visual substitution system for blind people based on SIFT description," *6th Int. Conf. Soft Comput. Pattern Recognition, SoCPaR 2014*, pp. 300–305, 2015.

[5] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, 2019.

[6] L. Sroba, J. Grman, and R. Ravas, "Impact of Gaussian noise and image filtering to detected corner points positions stability," *2017 11th Int. Conf. Meas. Meas. 2017 - Proc.*, vol. 10, no. 1, pp. 123–126, 2017.

[7] A. Khumaidi, E. M. Yuniarno, and M. H. Purnomo, "Welding defect classification based on convolution neural network (CNN) and Gaussian Kernel," *2017 Int. Semin. Intell. Technol. Its Appl. Strength. Link Between Univ. Res. Ind. to Support ASEAN Energy Sect. ISITIA 2017 - Proceeding*, vol. 2017-January, pp. 261–265, 2017.

[8] J. Nagi *et al.*, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," *2011 IEEE Int. Conf. Signal Image Process. Appl. ICSIPA 2011*, pp. 342–347, 2011.

[9] K. Selvakumar, S. Prabu, and L. Ramanathan, "Centroid neural network based clustering technique using competitive learning," *6th Int. Conf. Cond. Monit. Mach. Fail. Prev. Technol. 2009*, vol. 1, no. October, pp. 389–397, 2009.

[10] I. Marzuqi, G. P. Arinata, Z. M. A. Putra, and A. Khumaidi, "Segmentasi dan Estimasi Jarak Bola dengan Robot Menggunakan Stereo Vision," pp. 140–144, 2017.

[11] Syai'in, M., et al. *Smart-Meter based on current transient signal signature and constructive backpropagation method*. in *2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering*. 2014.

978-1-7281-2156-7/19/$31.00 ©2019 IEEE