# ADAPTIVE DEEP NEURAL NETWORK ALGORITHM FOR REAL TIME COLOUR VIDEO OBJECT DETECTION

TAN MEI YAN

SUPERVISOR
DR. ANIS FARIHAN BINTI MAT RAFFEI

FACULTY OF COMPUTING
UNIVERSITI MALAYSIA PAHANG

# 1 CHAPTER 1

## INTRODUCTION

### 1.1 Background of the problem

The last several years have witnessed the rapid development of scene understanding in the field of computer vision, especially the fundamental object detection task. The object detection task is to simultaneously localize the bounding boxes of objects and identify their categories in an image. There are three approaches to object detection which are deep learning, machine learning, and classical computer vision.

Convolutional neural network is used in deep learning approach to perform object detection while machine learning approach have 2 stages which is feature extraction and training classifier to perform object detection. The traditional computer vision training the classifier based on the target feature such as scale invariant feature transform (Lowe. David G, 2004). However, most of the traditional are using simple and easy method such as edge detection, motion estimation and optical flow and it require manual feature extraction and does not have high accuracy.

Video object detection extends this task to video sequences, which requires detectors to utilize multiple frames in a video to detect objects over time, which is another emerging topic in computer vision. Due to the complexity the similarity between each scene in of a video, it is hard to identify object. The appearance of objects might deteriorate significantly in some frames of a video, which could be caused by motion blur, video defocus, part occlusion, or rare poses. However, rich context information in temporal domain provides clues and opportunities to improve the performance of the object detection in videos.

## 1.2 Statement of the problem

As frames in a video clip are highly correlated, a larger quantity of video labels is needed to have good data variation, which are not always available as the labels are much more expensive to attain. It is easy to train an image object detector, but not always possible to train a video object detector if there are insufficient video labels for certain classes. There are more datasets for the image object detection task than for the video object detection task. It is much more expensive to collect labels for video datasets as there are more frames to label as the frames in a video clip are highly correlated. However, the existing image quality deterioration problem undermines its performance greatly, and various methods have been brought out to best utilize the video data to handle the quality deterioration problem.

## 1.3 Objective of the study

- To identify the limitation of existing real time colour video detection algorithms.
- To develop an adaptive deep neural network algorithm for real time colour video object detection.
- To evaluate the performance of the adaptive deep neural network algorithm for real time colour video object detection

# 2 CHAPTER 2

# LITERATURE REVIEW

## 2.1 Existing Related Works

Wearable Mobility Aid (WMA), VI Assistive System and Electronic Travel Aids (ETA) will be discussed.

WMA is a system that will notify people that suffering of visual impairment to receive their surrounding details using audio messages and vibro-tactile glove (M. Poggi, S. Mattoccia, 2016). This system is suitable for both indoor and outdoor surrounding. Tactile glove and RGBD camera are used to detect surrounding image. RANSAC algorithm will used to obtain ground plane information and CNN will train the system. The system is embedded using embedded CPU Torch 7. The system will send real time image and the user will receive the audio message using the application in mobile phone in 30ms. The system has a weight of 250g. The batter life is 3 hours for a 30 mA/h battery and 10 hours for a 10000 mA/h battery.

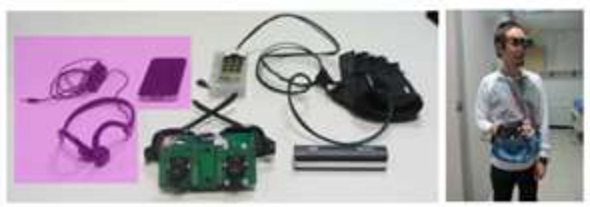


**Figure 1 Hardware Component of WMA**

[NAVIGATOR] Go straight ahead for 100 m and then turn right.
Warning: there is a trash-bin on the left at 2.2 m!

**Figure 2 Overview of WMA**

VI assistive system is a deep learning based assistive system to predict safe and reliable walkable instruction for visually impaired (Y. Lin, K. Wang, W. Yi, S. Lian, 2019). The system is suitable for both indoor and outdoor surrounding. RGBD camera are used to detect surrounding image. CNN is used in the system to train and predict collision-free instruction. The system is embedded using CPU Intel I7 8700K and GPU NVIDIA GeForce GTX1080. User can understand their surrounding through touchscreen interaction and earphone.  The system has a weight of 150g.

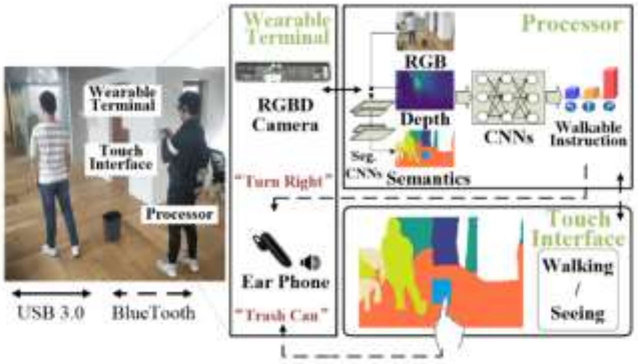


**Figure 3 VI Assistive System Overview**

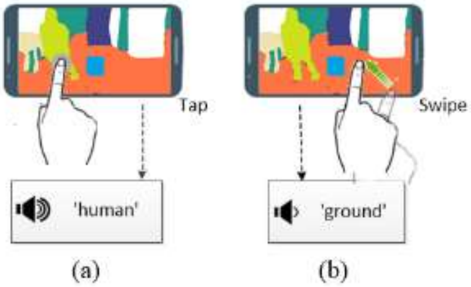


**Figure 4 Application of VI Assistive System**

ETA is a smart guiding device that give visually impaired guidance in indoor environment. RGBD camera, MCU and tactile sensor is used in this system to detect surrounding image. Depth-based way finding algorithm is used to find candidate traversable direction. The response time for the system is 30.2ms. The system can detect is from 0.4m to 4m (J. Bai, S. Lian, Z. Liu, K. Wang, Di. Liu , 2017).

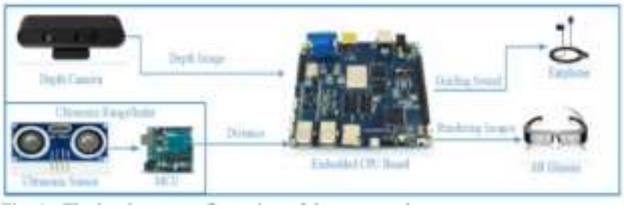


**Figure 5 Hardware Component of ETA**

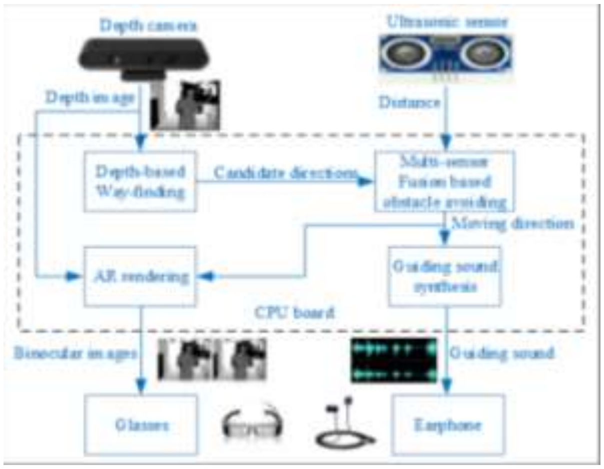


**Figure 6 Design of ETA**

## 2.2 Advantage and Disadvantage of Three Related Works

The advantages and disadvantages of WMA, VI Assistive System and ETA will be discussed.

A robust Random Sample Consensus (RANSAC) framework is used in WMA. The RANSAC algorithm proposed by Fischler and Bolles is a general parameter estimation approach designed to cope with a large proportion of outliers in the input data. RANSAC is a simple, yet powerful, technique that is commonly applied to the task of estimating the parameters of a model, using data that may be contaminated by outliers. WMA used RANSAC framework to detects the ground plane conditions, for example the presence of obstacles, holes, or stones. WMA also use a vibro-tactile glove to detect surrounding environment (J. Bai, S. Lian, Z. Liu, K. Wang, Di. Liu , 2017). There are three micro-motors placed in inch finger, middle finger, and pinky finger of the vibro-tactile glove. This glove is driven by the GPIO of the Odroid. This device will send feedback to user by vibrate and audio message. The major drawback of the feedback system is user cannot set the frequency of the audio message and vibration. When the distance between the user and obstacles is near, the feedback system will keep on vibrate or send audio message to user. Another drawback of WMA is the battery life. WMA can only work for 3 hours with a 3000 mA/h battery and 10 hours with a 10000 mA/h battery, so it is not suitable for user to travel long time.

The wearable terminal VI Assistive System consists of one sunglass, a stereo based RGBD camera and a Bluetooth earphone. This wearable terminal has a total weight of 150g. It is suitable for blind people wear and walk around easily. This system adopts an easy-to-use interaction design including touch screen and voice play. The user can then touch the screen and swipe around with a single finger to get the information of their surrounding environment. This VI Assistive System have two major drawback which are perspective view and performance. When the distance of user and object is too close, the perspective view will be lacked and difficult to detect surrounding environment. The night performance of VI Assistive System is low. Even though the system can work at night, but the system may give a wrong instruction (Y. Lin, K. Wang, W. Yi, S. Lian, 2019).

The maximum cost for ETA to process every frame is 30.2ms. The time cost to detect unfamiliar environment is better than cane. The user can get feedback from the system in real time, so the obstacle can be detected timely. The ultrasonic sensor is mounted on the ETA glasses. The ultrasonic rangefinder can obtain the distance of object and user by sending and receiving wave. There are two main drawbacks in ETA, which are environment and user experience. The main drawback of the system is it cannot work at outdoor environment (J. Bai, S. Lian, Z. Liu, K. Wang, Di. Liu , 2017). Based on user's experience, the recorded instructions-based ETA is less efficient as they fell that it is hard to turn the accurate angle.

### 2.3 Comparison Between R-CNN, R-FCN and YOLO

There are many different types of algorithms used in object detection. This chapter will discuss the R-CNN (Region-based Convolutional Neural Networks), R-FCN (Region-based Fully Convolutional Network), and YOLO (You Only Look Once)

#### 2.3.1 R-CNN

R-CNN will divide the image into multiples boxes call as bounding boxes. Then the bounding boxes will group by different sizes, texture, color and intensity by using different algorithm. The similarity of the color in the image can be measured by color histograms intersection using RGB (J. R. R. Uijlings,K. E. A. van de Sande, T. Gevers,A. W. M. Smeulders, 2018).

$$S_{color}(r_i , r_j) = \sum_{k=1}^{n} \min\,(c_i^k , c_j^k)$$

SIFT measurement is used to measure the texture similarity in the image. (J. R. R. Uijlings,K. E. A. van de Sande, T. Gevers,A. W. M. Smeulders, 2018)

$$S_{texture}(r_i , r_j) = \sum_{k=1}^{n} \min\,(t, t_j^k)$$

Size similarity can help small bounding box merge easily.

$$S_{size}(r_i\, ,r_j) = 1 - \frac{size(r_i) + size(\, r_j)}{size(im)}$$

Fill similarity can measure the fitness of the bounding box. Thus, bounding box with higher will merge together first to avoid holes. The fill similarity can calculate by using following algorithm.

$$fill(r_i\, ,r_j) = 1 - \frac{size(BB_{ij}) - size(r_i) - size(\, r_j)}{size(im)}$$

The final similarity can measure by combining the four algorithms above where $a_1 \in \{0,1\}$.

$$s(r_i\, ,r_j) = a_1\, S_{color}(r_i\, ,r_j) + a_2\, S_{texture}(r_i\, ,r_j) + a_3\, S_{size}(r_i\, ,r_j) + a_4\, S_{fill}(r_i\, ,r_j)$$

After the final similarity is obtained, the bounding box are then group into different squares. The squares will train and the object in the square can classified by using CNN one by one. Figure below shows the general process of R-CNN.

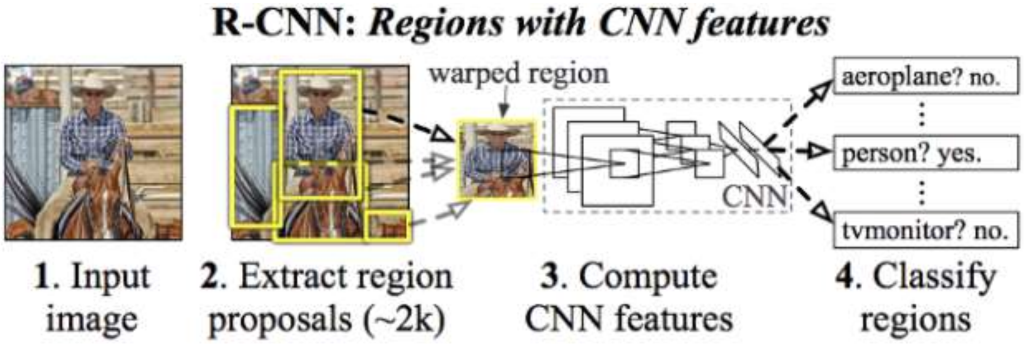


**Figure 7 R-CNN**

### 2.3.2 R-FCN

R-FCN is a combination of Region Proposal Network (RPN) and Fully Convolutional Network (FCN). In the process of calculation, RPN and FCN have shared convolution layer, realize the weights of sharing, reduce the number of weights, can greatly reduce the number of parameters in the architecture and reduces the complexity of the network model. At the same time, the consistency of feature mapping in RPN and FCN is guaranteed, and the positioning accuracy is improved.

RPN Figure below shows then structure of R-FCN and the data flow for R-FCN.

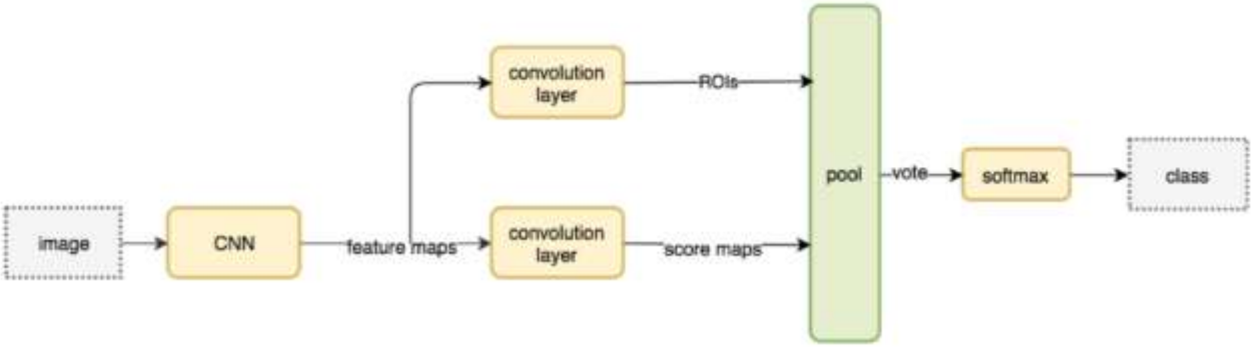


**Figure 8 Data Flow of R-FCN**

Before to pool, the square inside the image will divide into 3*3 regions, which is also call as 9 feature maps that contains some parts of the object. The feature maps shared between RPN and R-FCN are computed. Average score for each feature map will calculate based on the percentage of the object covered in the feature map. The average score for each part will calculate in the pool. The score map can be calculated by using position-sensitive ROI pooling operation (J. Dai, Y. Li, J. Sun, 2016).

$$r_c(i_j\,, \Theta) = \sum_{(x,y)\epsilon\, bin(i,j)} z_{i,j,c(x+x_0\,,y+y_0|\Theta)/n}$$

### 2.3.3 YOLO (You Only Look Once)

YOLO is one of the famous object detection technique used in real time video. In YOLO, the image will be divided into S*S grid which call as cells. Each of the cell can create 5 bounding boxes that intersection with other cells. The bounding boxes is a rectangle box than will include object inside the box. 5 predictions which are x, y, w, h and confidence score will be done in each bounding box. (x, y) represent the coordinates of the box and (w, h ) represent the width and height of the bounding box. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts (J. Redmon, S. Divvala, R. Girshick, A. Farhadi, 2016). The confidence score can measure by using the following algorithm.

$$\mathrm{Pr(Object)} \ * \ IOU_{pred}^{truth}$$

If no object inside the bounding box, the confidence score is equal to zero. Otherwise confidence score to equal the intersection over union (IOU) between the predicted box and the ground truth (J. Redmon, S. Divvala, R. Girshick, A. Farhadi, 2016). PASCAL VOC dataset is used to train YOLO object. It can identified up to 20 different classes of object The class probabilities can measure by using following algorithm.

$$\mathrm{Pr(Class|Object)} * \mathrm{Pr(Object)} \ * \ IOU_{pred}^{truth} = \mathrm{Pr}(Class_i) * \ IOU_{pred}^{truth}$$

Each bounding box will have their own class-specific scores. These scores encode both the probability of that class appearing in the box and how well the predicted box fits the object. At last, the box with confidence score more than 30% will keep and the others box with confidence score less than 30% will be removed. Figure below shows the process of YOLO.

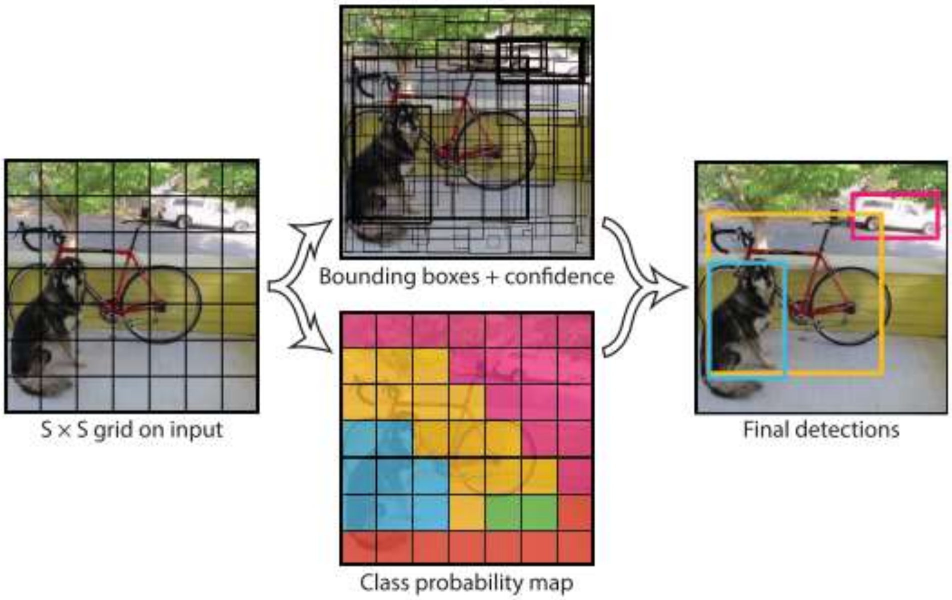


**Figure 9 General Process of YOLO**

# 3 CHAPTER 3

# RESEARCH METHODOLOGY

## 3.1 Dataset

There are many datasets that can used for object detection. The dataset that used in this system is Microsoft's Common Objects in Context (COCO) dataset. COCO dataset is one of the famous datasets that used in object detection, segmentation and captioning dataset. There is total or 330K images including 200K + annotated, and more than 2 million of instances that categorize in 80 object categories. There are 250,000 of people dataset in COCO dataset.

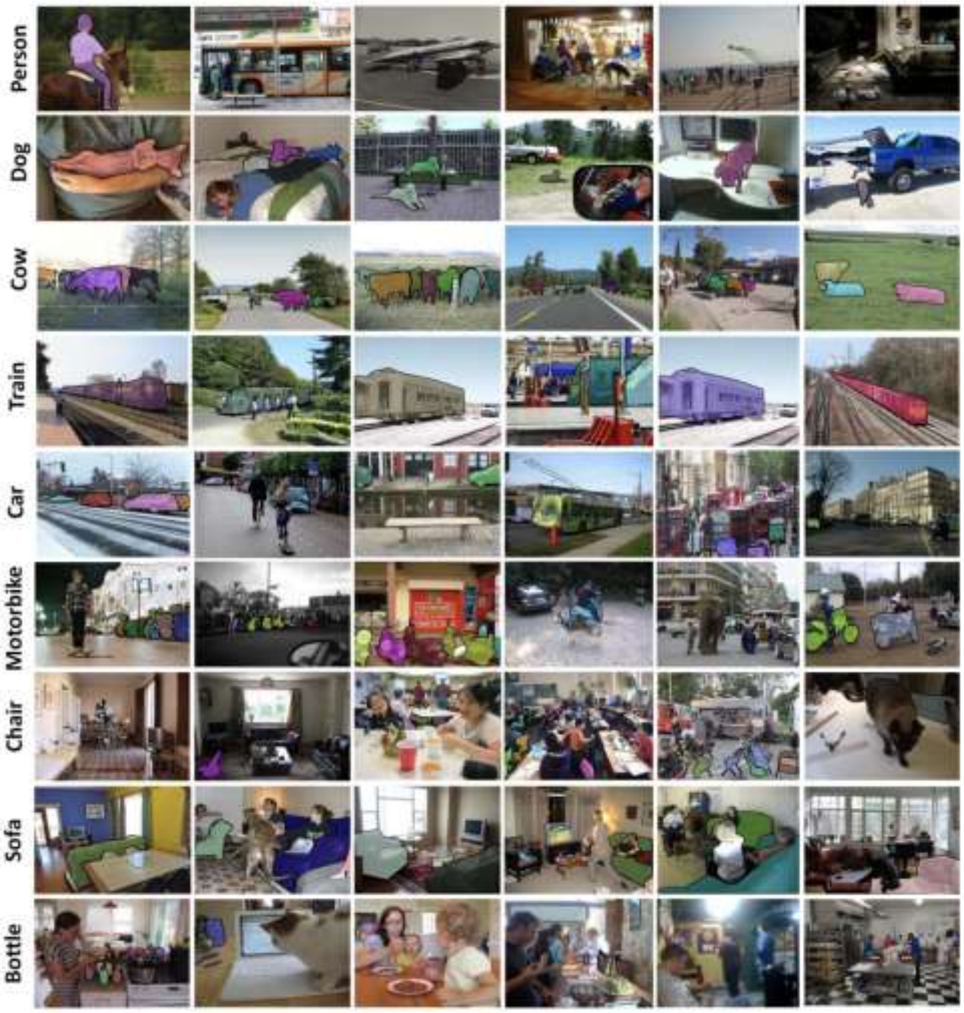


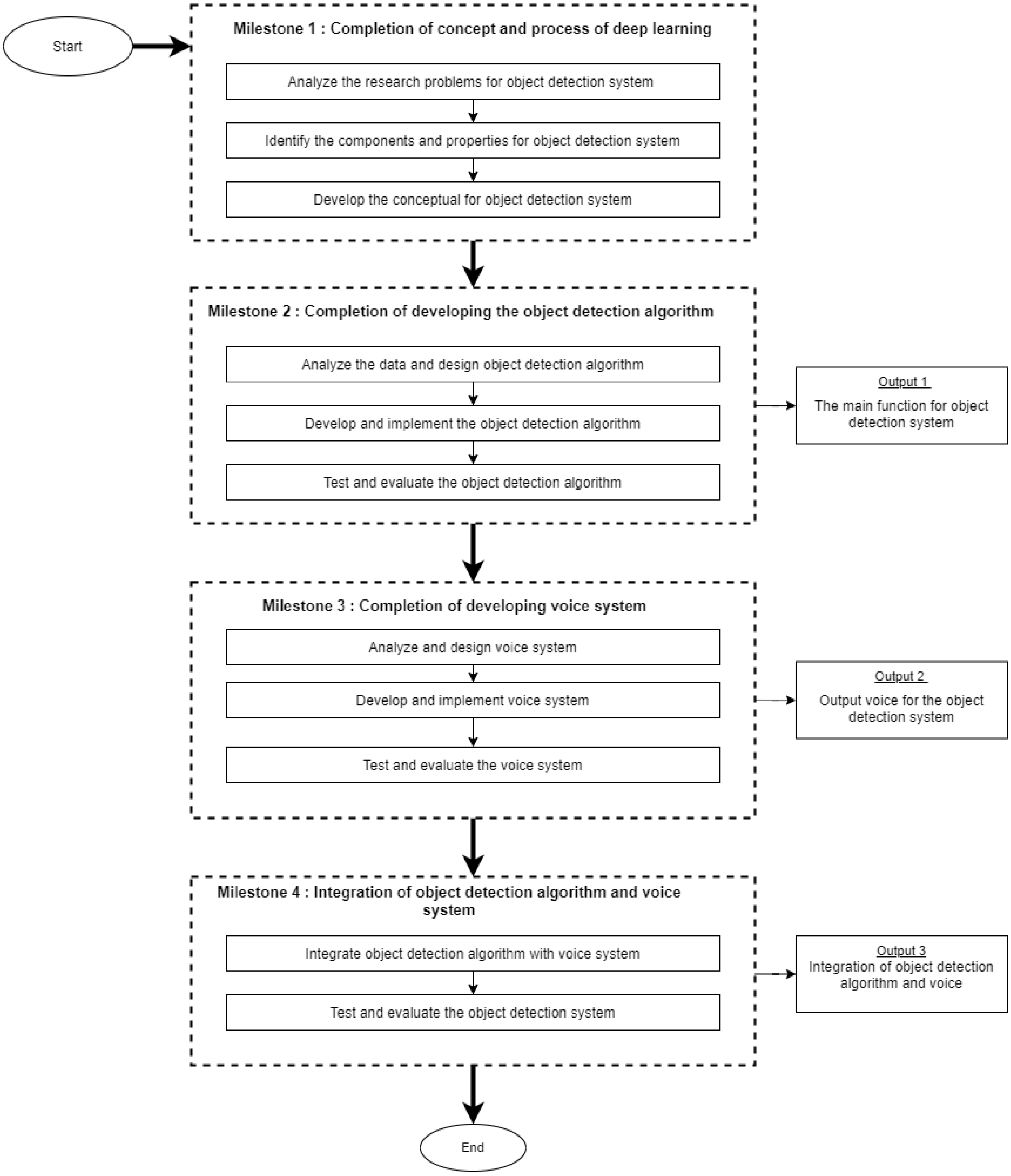
**Figure 10 Example of COCO dataset**

# 4 Research Framework

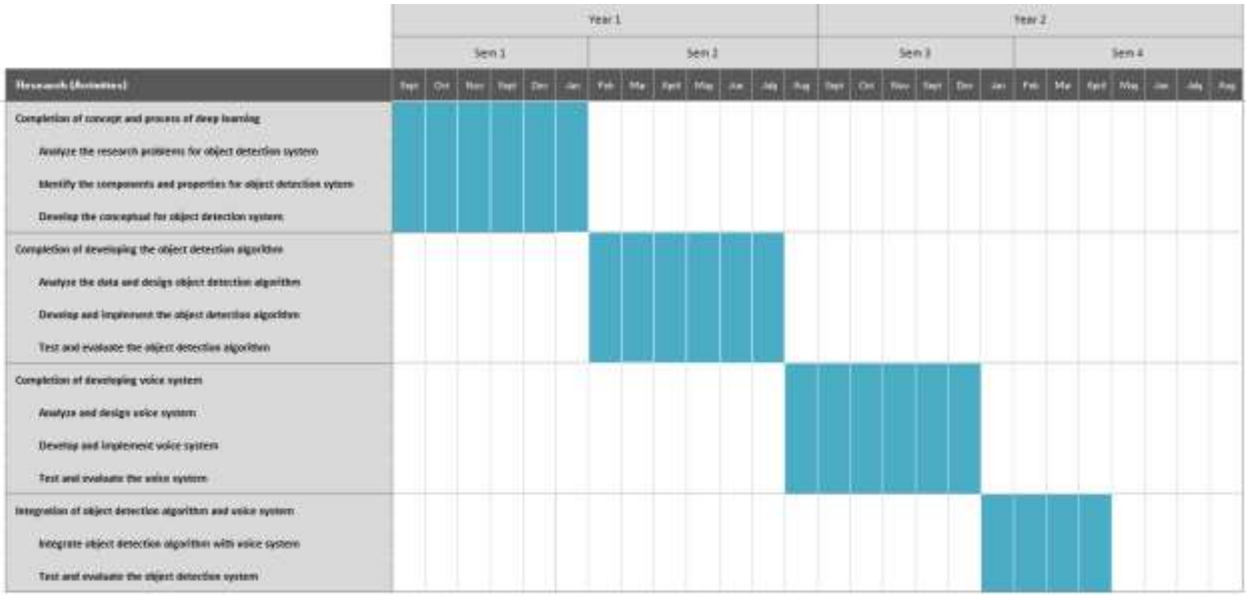
Refer to the figure of operational framework below.

## 3.3 Research Framework

Refer to the figure of operational framework below.


Start
Milestone 1 : Completion of concept and process of deep learning
Analyze the research problems for object detection system
Identify the components and properties for object detection system
Develop the conceptual for object detection system
Milestone 2 : Completion of developing the object detection algorithm
Analyze the data and design object detection algorithm
Develop and implement the object detection algorithm
Test and evaluate the object detection algorithm
Output 1
The main function for object detection system
Milestone 3 : Completion of developing voice system
Analyze and design voice system
Develop and implement voice system
Test and evaluate the voice system
Output 2
Output voice for the object detection system
Milestone 4 : Integration of object detection algorithm and voice system
Integrate object detection algorithm with voice system
Test and evaluate the object detection system
Output 3
Integration of object detection algorithm and voice
End

# Gantt Chart



| | Year 1 | | | | | | | | | | | | | Year 2 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sem 1 | | | | | | Sem 2 | | | | | | | Sem 3 | | | | | | Sem 4 | | | | | | |
| Research (Activities) | Sept | Oct | Nov | Sept | Dec | Jan | Feb | Ma | April | May | Jun | July | Aug | Sept | Oct | Nov | Sept | Dec | Jan | Feb | Mar | April | May | Jun | July | Aug |
| Completion of concept and process of deep learning | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | | | | | | | |
| Analyze the research problems for object detection system | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | | | | | | | |
| Identify the components and properties for object detection sytem | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | | | | | | | |
| Develop the conceptual for object detection system | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | | | | | | | |
| Completion of developing the object detection algorithm | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | |
| Analyze the data and design object detection algorithm | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | |
| Develop and implement the object detection algorithm | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | |
| Test and evaluate the object detection algorithm | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | |
| Completion of developing voice system | | | | | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | |
| Analyze and design voice system | | | | | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | |
| Develop and implement voice system | | | | | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | |
| Test and evaluate the voice system | | | | | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | |
| Integration of object detection algorithm and voice system | | | | | | | | | | | | | | | | | | | ■ | ■ | ■ | ■ | | | | |
| Integrate object detection algorithm with voice system | | | | | | | | | | | | | | | | | | | ■ | ■ | ■ | ■ | | | | |
| Test and evaluate the object detection system | | | | | | | | | | | | | | | | | | | ■ | ■ | ■ | ■ | | | | |

# 5 References

J. Bai, S. Lian, Z. Liu, K. Wang, Di. Liu . (2017). Smart Guiding Glasses for Visually Impaired People in Indoor Environment. *IEEE Tranation of Consumer*, 258-266.

M. Poggi, S. Mattoccia. (2016). A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning. *IEEE Computer Community*, 208-213.

Y. Lin, K. Wang, W. Yi, S. Lian. (2019). Deep Learning based Wearable Assistive System for Visually Impaired People.