T.C. İSTANBUL TİCARET ÜNİVERSİTESİ

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

A MACHINE LEARNING BASED VIDEO PROCESSING WARNING SYSTEM FOR LIFEGUARDS

Zahraa Qosay Hashim ALHAIDARI

Supervisor
Assist. Prof. Dr. Mustafa Alper ÖZPINAR

MASTER'S THESIS

DEPARTMENT OF MECHATRONICS ENGINEERING

ISTANBUL - 2019

ACCEPTANCE AND APPROVAL PAGE

On 27/06/2019, Zahra Qosay Hashim Alhaidari successfully defended the thesis, entitled **"A Machine Learning Based Video Processing Warning System for Lifeguards"**, which she prepared after fulfilling the requirements specified in the associated legislation, before the jury members whose signatures are listed below. This thesis is accepted as a **Master's Thesis** by Istanbul Commerce University, Graduate School of Natural and Applied Sciences **Mechatronics Engineering Department.**

Approved by:

Supervisor — **Asst. Prof. Dr. Alper ÖZPINAR**
İstanbul Commerce University ....

Jury Member — **Asst. Prof. Dr. Muhammet Ceylan**
İstanbul Commerce University ....

Jury Member — **Asst. Prof. Dr. Erdinç ÖZTÜRK**
Sabancı University ....

Onay Tarihi : 23.10.2019

Prof. Dr. Necip ŞİMŞEK
Enstitü Müdürü

DECLARATION OF ACADEMIC AND ETHIC INTEGRITY

I hereby declare that

• I have obtained all information and documents within the framework of academic rules,

• I present all visual, audio and written information and results in accordance with scientific ethics,

• I refer to the relevant works in accordance with scientific norms if others' works are used and I cite all the work I refer to as reference,

• and that I have not submitted any part of this thesis as another dissertation at this university or another university.

Date

24/05/2019

Signature

Name and Surname

Zahraa Qosay Hashim ALHAIDARI

TABLE OF CONTENTS

ABSTRACT

M.Sc. Thesis

A MACHINE LEARNING BASED VIDEO PROCESSING WARNING SYSTEM FOR LIFEGUARDS

Zahraa Qosay Hashim ALHAIDARI

Istanbul Commerce University
Graduate School of Natural and Applied Sciences
Department of Mechatronics Engineering

Supervisor: Assist. Prof. Dr. Mustafa Alper ÖZPINAR

2019, 60 pages

In various real-life applications Object detection in videos plays an important role. the classic approaches that depend on use handcrafted features which are optimal for specific tasks and problem-dependent. Moreover, the handcrafted features are highly sensitive to dynamical events such as camera jitter, illumination changes, and changing in object sizes. On the other hand, because increasing of crimes and mishaps as well as car accidents there is a need of continuous surveillance to keep an eye on public areas such as historical places, beaches, air ports. For automate monitoring the object in a video file many algorithms and technology have been developed. One of challenging tasks when dealing with computer vision is Object detection and tracking in videos and images. Mainly the steps for analyzing video are: First detection the objects of interest from other objects in motion, Stat tracking the interested objects in consecutive frames, after detection and tracking the final step is Analyzing the behavior of that object.

This paper represents an automated machine learning based video processing for detecting and tracking the people at beaches and pools and works an expert warning system for lifeguards. The real-world example based on tracking by using RFID triggered tracking or zone-based tracking where people either presents themselves at the tracking point by RFID bracelets or staying in the labelling zone for limited amount of time.

Keywords: Classification, image processing, lifeguard system, machine learning, object tracking, video processing.

ÖZET

Yüksek Lisans Tezi

CANKURTARANLAR İÇİN MAKİNE ÖĞRENİMİ TABANLI VİDEO İŞLEME UYARI SİSTEMİ

Zahraa Qosay Hashim ALHAIDARI

İstanbul Ticaret Üniversitesi

Fen Bilimleri Enstitüsü

Mekatronik Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Mustafa Alper ÖZPINAR

2019, 60sayfa

Çeşitli gerçek hayattaki uygulamalarda Videolarda nesne algılama önemli bir rol oynar. bağımlı olan klasik yaklaşımlar, belirli işler için en uygun ve probleme bağlı el yapımı özellikleri kullanır. Dahası, el yapımı özellikler kamera titremesi, aydınlatma değişiklikleri ve nesne boyutlarında değişiklik gibi dinamik olaylara karşı oldukça hassastır. Öte yandan, suçların ve aksiliklerin yanı sıra araba kazalarının artması nedeniyle, tarihî yerler, plajlar, hava limanları gibi kamuya açık alanlara göz atmak için sürekli bir gözetim ihtiyacı vardır. algoritmalar ve teknoloji geliştirilmiştir. Bilgisayar vizyonuyla uğraşırken zorlu görevlerden biri, video ve görüntülerde Nesne algılama ve izlemedir. Temel olarak videoyu analiz etme adımları şunlardır: Önce hareket halindeki diğer nesnelerden ilgilenilen nesneleri tespit etme, İlgili nesneleri ardışık karelerde izleme, son adımı tespit etme ve izleme sonrasında bu nesnenin davranışını analiz etme yöntemidir.

Bu makale, plajlardaki ve havuzlardaki insanları tespit etmek ve izlemek için kullanılan otomatik bir makine öğrenmesi tabanlı video işlemeyi temsil ediyor ve cankurtaranlar için uzman bir uyarı sistemi kullanıyor. RFID tetiklemeli izleme veya insanların izleme noktasında RFID bilezikleri tarafından sunulan veya etiketleme alanında sınırlı bir süre kaldıkları bölgeye göre izleme kullanarak izlemeye dayalı gerçek dünya örneği.

Anahtar Kelimeler: Cankurtaran sistemi, görüntü işleme, makine öğrenmesi, nesne izleme, video işleme, sınıflandırma.

ACKNOWLEDGEMENTS

I would like to express my thankfulness to Assist. Prof. Dr. M. Alper ÖZPINAR for his guidance, suggestions and advice throughout the complete of this thesis.

Finally, I would like to express my gratitude to my family, for all their support and love.

Zahraa Qosay Hashim ALHAIDARI

İSTANBUL, 2019

LIST OF FIGURES

LIST OF TABLES

SYMBOLS AND ABBREVIATIONS LIST

SUV	Sport-utility vehicle
RGB	Red, green, and blue
MIT	Massachusetts Institute of Technology
LID	Light-emitting diodes
PLC	Programmable logic controller
CMOS	Complementary metal-oxide-semiconductor
HOG	Histogram of Oriented Gradients
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
PASCAL VOC	PASCAL Visual Object Classification
COCO	Common Objects in Context
IOU	Intersection over Union
MAP	Mean Average Precision
NMS	Non-Maximum Suppression
CNN	Convolutional Neural Network
CNN-R	Region-based Convolutional Network
ROI	Regions of Interest
Fast R-CNN	Fast Region-based Convolutional Network
RPN	Region Proposal Network
YOLO	You Only Look Once
SSD	Single-Shot Detector
MOT	Multi Object Tracking

1. INTRODUCTION

Computers have become an important tool in different lives field. They execute computational mission very carefully, and extra efficiently than the humans. To carry out more difficult tasks such as analysis of visual scenes or speech, logical inference and reasoning, capabilities of computer expand day after day, the high-level duty that the humans implement subconsciously several times every day with numerous eases that human do not commonly even realis that they implementing them. (Navneet, 2006).

One of the fundamental tools is analyzing and describe object in images and videos for a wide of applications such as personal photo collection management, virtual personal shopping assistant, surveillance, contextual advertisement and human computer interaction. (Dantone, 2014). Moreover smart-phones cameras are Cheap and small, with high quality, advances in camera systems led computer devices and social networks to create a high need for computer vision algorithms that qualify an automatic classification and characterization the object in images and videos. (Rajkamal, 2014).

The increase in crimes lead to huge demand on the security systems, video surveillance is one of the large research lines. The people critical requirement for providing security to themselves and people properties lead to huge development and amplification in video surveillance systems around the world. (Martin, 2009). The computer vision, digital processing for image and video research contains set of algorithms of motion detection, tracking object, object detection, event detection, etc., these algorithms are used in security. One of the complexities in different video surveillance applications the difficult of modeling a person and his activities because the big variability in physical appearance, movements and interactions between the people and the object. Object detection and tracking algorithms includes various solutions for solving this problem.

The object detection processing phase of analysis automated video system involve: background/foreground extraction, object extraction (object detection), object classification, object tracking, and event or action recognition. (Valera and Velastin, 2005) and (Hu, et al., 2004). As shown in Figure 1.1 system contain of four phases:

- Background extraction: Motion detection is the first step in each visual surveillance approach. The purpose of motion detection at segmenting area matching to moving objects from the remainder of the image. The sequential stages build on the background accuracy gain, that is, the remnant of stages have a powerful dependency with results obtained in this operation: false object detections, partial object detections or missing objects could be happened because of bad background model. (Piccardi, 2014).

- Object extraction: the second step is applying morphological operations to improve the object extraction and decrease the noise in the produced image mask. After detecting the boundaries of the region, it is also useful to get rid of regions which the boundary does not separated them. A linked component analysis is realized so as to segment or extract objects. In the following processes will analyze the objects that extracted or detected in this step only.

-Object tracking: Surveillance systems start track moving objects, after motion detection and object extraction. By locating the object placement in each frame of a video sequence, the object tracker target is to create the trajectory of an object over time. Tracking process include matching the objects in consecutive frames by utilize different features like points, lines or blobs. Tracking methods can be divided into three major categories: point, kernel or silhouette tracking. (Yilmaz, et al., 2006).

-Object classification: Object classification process match object classification models to generate object models from an image or sequence (tracking) and the final decision will be made according to their similarity.

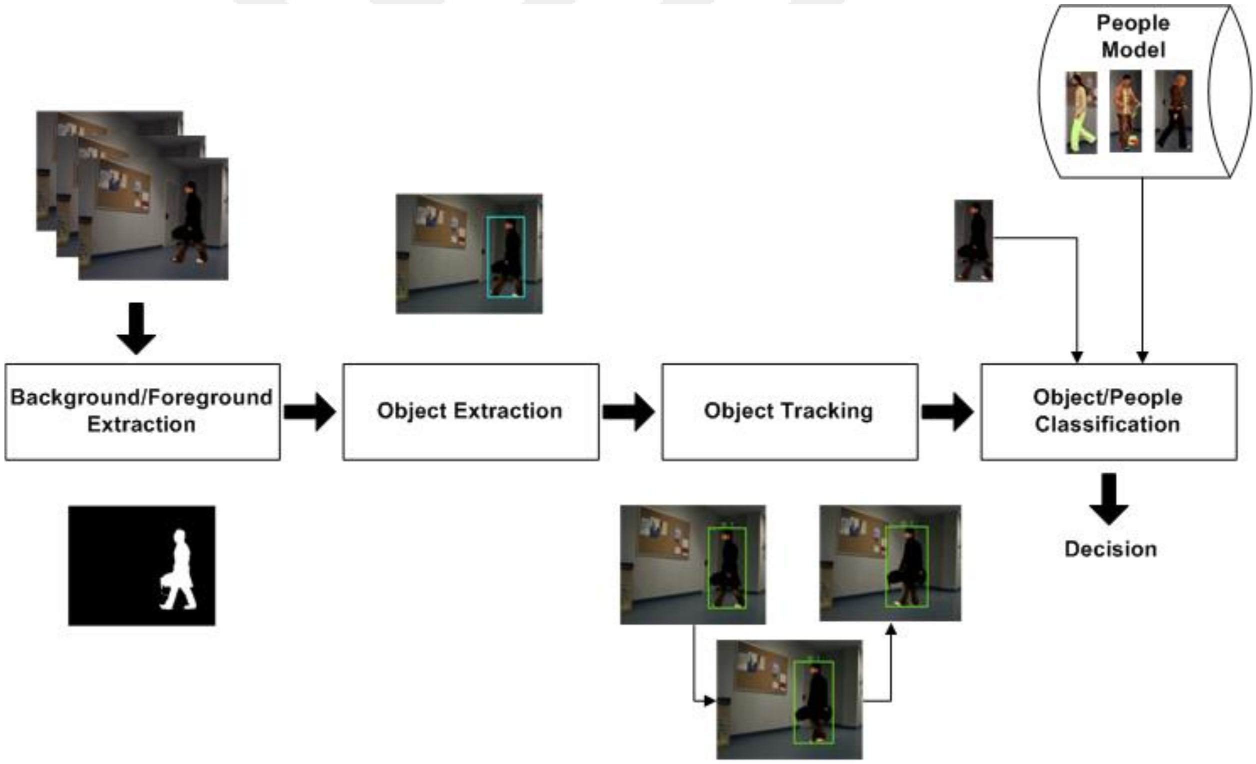


Figure 1.1. System comprises of four stages

According to Kloss (2015), visual perception is one of the most significant senses, for humans and many other animals. For interacting with our environment, they rely on vision heavily whenever they want to: pick up an object, or when we recognize other people faces. For doing these tasks, they use object recognition and localization algorithms.in order to recognize human face first they need to find which part in the image we see represent the human face and where is the face. Human brain responsible for all this processing steps. For building computer algorithms for face recognition, this algorithm most tackle a small subset for different tasks which is necessary for understanding the image. In order to reproduce a part of human brain visual perception which seem effortless, artificial systems need to combine several different algorithms.

2. LITERATURE REVIEW OF IMAGE PROCESSING

Image processing can be defined as the process of using computer algorithms to convert an image into digital form, in order to improve its quality and implement some operations on the image and get a better image or to extract useful information from the image. (Jalled,2016). Usually There are two kind of operation can be used for image processing:

- Analogue image processing used for the hard copies, such as photographs and printouts. For using visual techniques image analysts employ several fundamentals of interpretation.
- Digital image processing techniques by using computers can assist in the digital image manipulation.

Image Processing system contain three basic steps:

- By digital photography and using optical scanner Importing the image and the image acquisition tools.
- Analyzing and manipulating the input image that cover data compression and image enhancement.
- The final process is the output in this stage the result will be changed image or statement according to the image Analysis.

Figure 2.1 shows the block diagram where the output image from processing operation will be used for object detection algorithms.

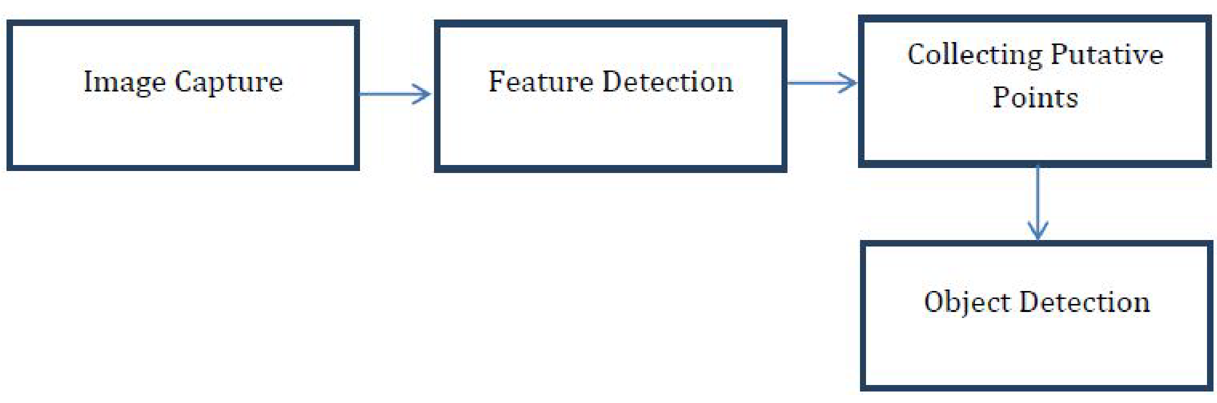


Figure 2.1. Block Diagram for Object Detection

Image processing different techniques dealing with the input image such a two-dimensional signal and applying standard techniques of signal processing to the image. A pixel, represent one of the picture elements, this pixel is collected by the camera's sensor can be count like a tiny dot containing information about the picture. This information stored in 3 planes. Each of these planes represent three colors which are Red, Green and Blue plane as shown in Figure 2.2. Each one of plane has 8-bit of information where the intensity from 0 up to 255 or per plane. There are three color in each plane the combination of these color makes an RGB image.

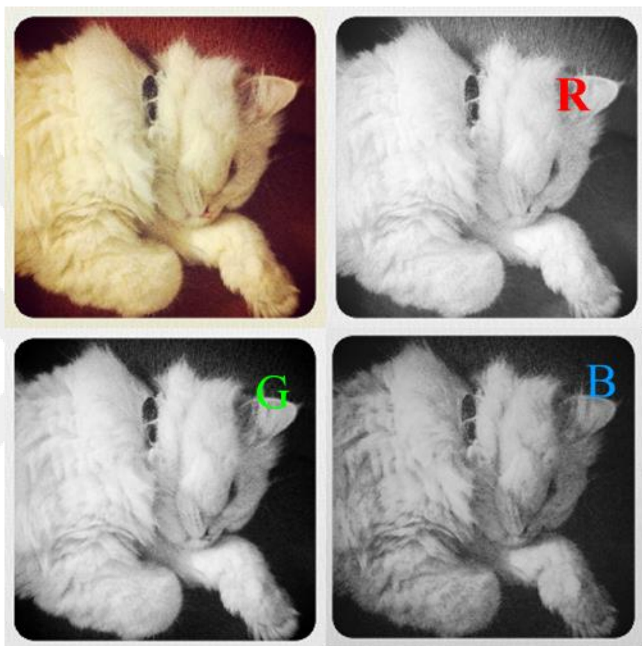


Figure 2.2.RGB image.

The purpose of digital image processing producing continuous image. This continuous representation is processing the outputs. The digital images will be used in the machine vision, and also in image analysis. There are two possible way for processing images from video, first one in form of video stream or another way as a separate picture. The important parameters in processing a digital image, the sampling interval and the quantization interval. (Kawahito and Seo, 2017). According to Blackledge (2009), the dimension is the basic variation between the signal processing and image processing. Digital image can be performed by two-dimensional function *F* (x, y), which is characterize by a matrix, where *x* performs rows and *y* perform columns. The size of the matrix relies on the upper image resolution. The most utilized image processing procedure are image filtration operations. Where image filtration stands for operations of image modification or enhancing an image, the output of image information with wanted quality. Image filtering processing operations include smoothing, sharpening, and edge enhancement. The next image information analysis is segmentation of image. The process of segmentation divide image to several objects. The purpose of segmentation is to divide or partition the image into various parts called segments, i.e. characteristic a background, because the background does not contain any information for processing. The threshold method stands for the operation that converts the image in gray-scale form into a binary image. (Gonzales and Woods, 2008).

To get the information from digital image, the image has to subjected various steps of processing, the flow chart in Figure 2.3 show this phase.

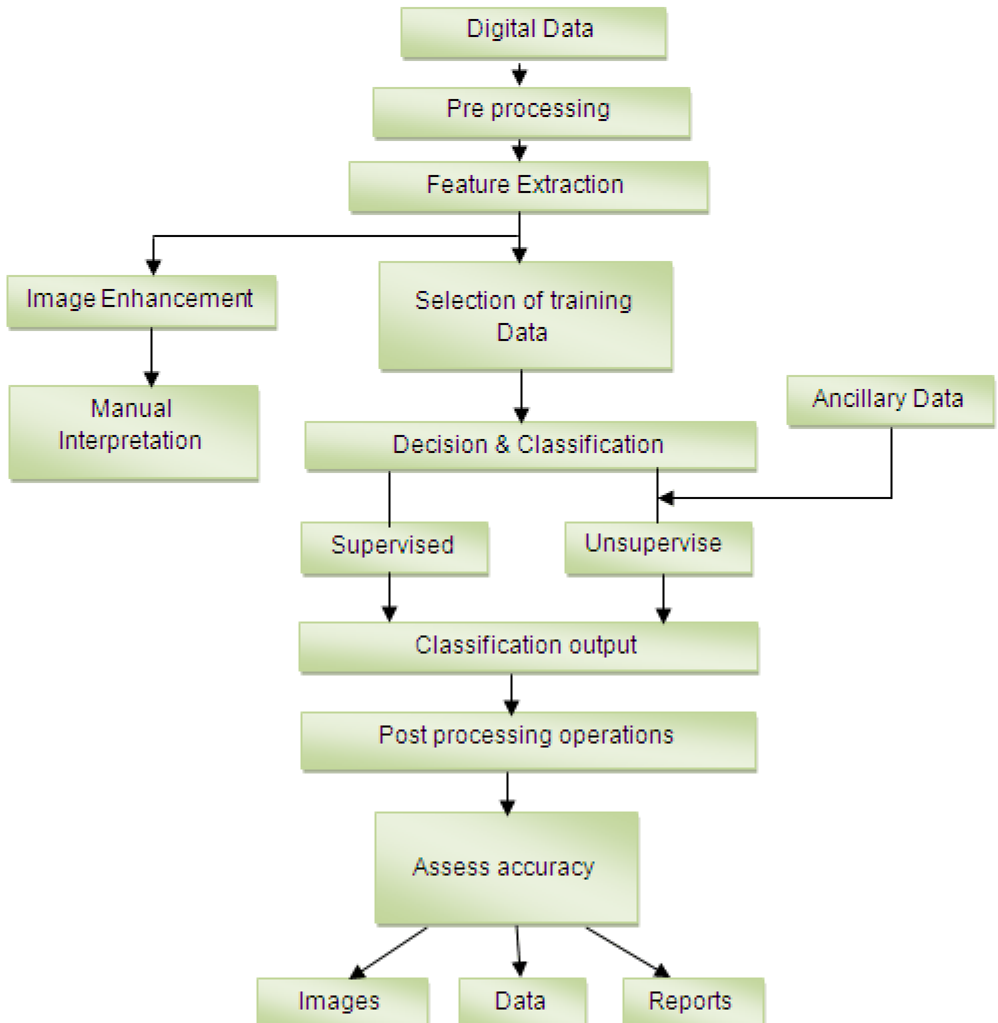


Figure 2.3. Flow diagram showing different steps that digital image subjected to

2.1. Flow Chart Design

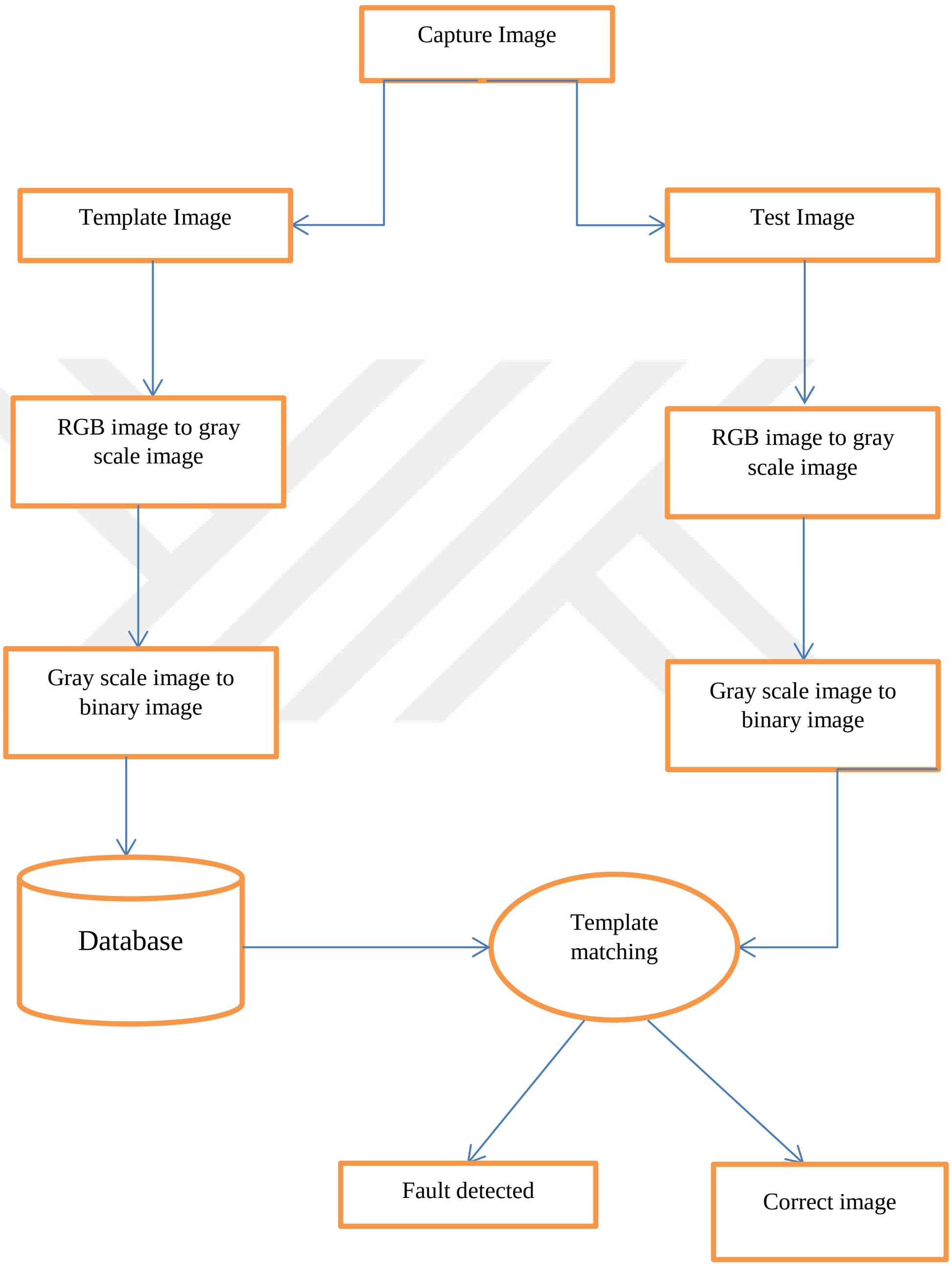


Figure 2.4.System Architecture

Figure 2.4. Shows the flow chart of Fault Detection System. After putting the product under the high-resolution camera so as to obtain an RGB image of the product. In order to extraction the information from RGB image we first need to converted to the form of gray scale. This gray scale image is the input for the binarization. Gray scale image is then converted to binary form. The template image is stored in the database. Every test image is match with this template image. Template matching is done using XOR operation. Results of template matching is stored in the database.

2.2. Image Conversion

The captured image by the camera is an RGB format image, for using this image in different process, the RGB will be converted to the binary format image. the grayscale form of the digital image is an image that the amount of each pixel is an elementary sample, this sample holds only intensity information. This kind of Images, can be define as black-and-white, which are contain only a shade of gray, where the smallest intensity value represents by black, while the higher one represents with white. (Johnson, 2006). First, the grayscale image will be converting to a binary image. In order to do that we will use thresholding technique. Thresholding replaces whole the pixels existing in the input image that belong to a gray-level interval level with the amount 1 (white) and change the whole other pixels with the amount 0 (black). When the input image is not in form of a grayscale image, then this image should be converted to grayscale image.

2.3. Template Matching

Template matching can be defined as a digital image processing technique, used for searching and finding tiny parts location in the image, these parts most be match a template image. One of the Template matching application is using this technique as a part of quality control in manufacturing. to verify the actual shape and dimension of an ideal product, the technic will compare every pixel inside the input binary form image with the template image. Where the template matching process will use the intensity amount of both the input image and the template image for making the comparison process. Template matching is a very common approach to detecting objects and finding the similarity measurement. Template matching tries to find a certain object in the input image and find the location of this object. In order to do this template matching compute a different measure between the template and input image. the product will be count as a complete piece when the template match carefully. Template matching requires two images to find areas of the image that look like a template image. Source image (I): The input image in which will be matched to the template image. Template image (T): The patch image that used to compared to the template image. (Hashemia, et al., 2016).

2.4. Applications of Template Matching

Template matching algorithms contribute in various applications, from security-based to biomedical-based projects. Although the application of this technique is restricted because of some limitations in shift and scale invariant feature matching, template matching remains a strong tool for extracting patterns in the pre- and post-processing steps of image analysis. Fast template matching and some deformable-based matching algorithms provide quick solutions for object recognition, while algorithm complexity is not very high.

2.4.1. Face detection

Face detection is a technique that is used to find an arbitrary sub-image representing the human face of a given and global image. This field of computer vision and image processing is broadly used in security applications, video surveillance, tracking, etc., and it may represent all or only a part of such systems. The traditional method of extracting a given face from a global image is to apply a matching algorithm to the target image. One of the most important factors in face detection is a set of human facial features. Over the past decade, a high volume of research has been performed in the domain of face detection, which resulted in the determination that the eyes, for face detection and recognition the most important features are mouth and nose. The facial feature extraction methods, this method is critical to various non-idealities also use the template matching methods, such as variations in illumination, noise, orientation, time consumption and color space used.

2.4.2. Eye detection

Eye detection is a precondition for numerous implementations, like iris recognition, human–computer interfaces, security, biology systems and driver drowsiness detection. The template is correlated with different areas of the facial image. The eye region is the region of face which provides the maximum correlation with the template. The template matching can be used effectively in closed eyes both as well as open eyes his can be seen in Figure 2.5.

Figure 2.5. Face images, in two phases closed and open eyes

2.4.3. Automatic bare PCB board testing

Another traditional application of template matching is in the detection of loose connections in electrical boards as shown in Figure 2.6. A bare printed circuit board is a PCB without any placement of electronic components, and it is utilized with other components to produce electric instruments at the circuit level. In order to reduce spending in manufacturing caused by a defective bare PCB, the bare PCB must be inspected.

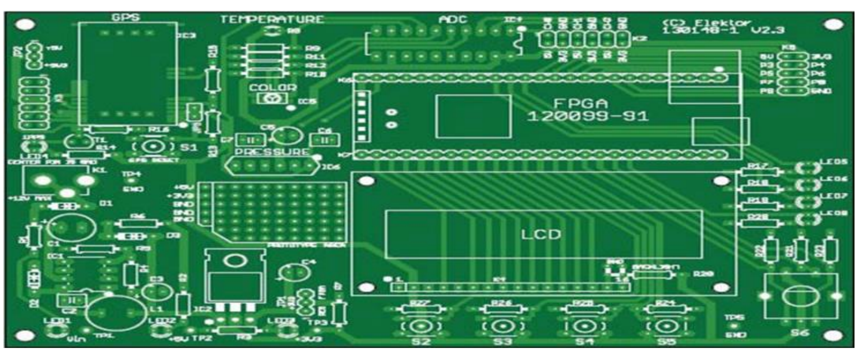


Figure 2.6. Bare PCB of Multi input/output

2.4.4. Normal breast and mammogram positioning

According to Sulaiman et al. (2007), they proposed image processing algorithm, this algorithm able to determine and locate the symmetrical value between breast mammograms and template matching as shown in Figure 2.7. They used template matching technic to implemented a new algorithm by using a cross-correlation method. Their 2D cross-correlation algorithm produced a highly accurate result in the matching process. For two-dimensional images cross-correlation algorithm operates well and get the best result for the matching process. the algorithm work for detecting similarity in grayscale mammograms. This algorithm helps to increase the breast cancer detection by 80%, this help to improve computer aided diagnosis system.

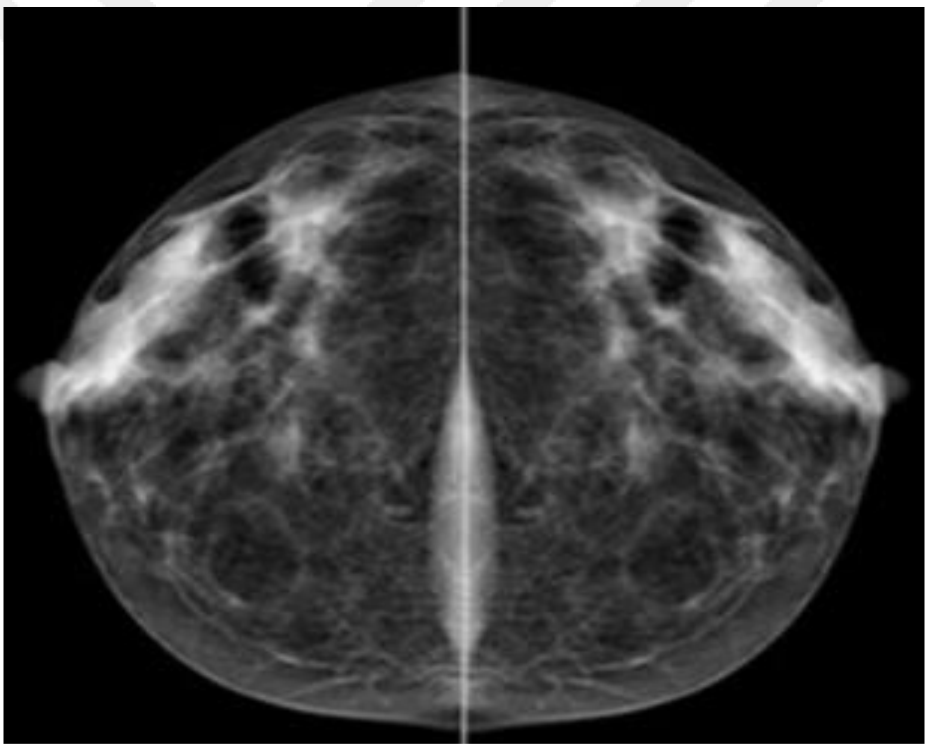


Figure 2.7. Symmetrical breast mammogram

2.5. Categorization of Computer Vision Tasks

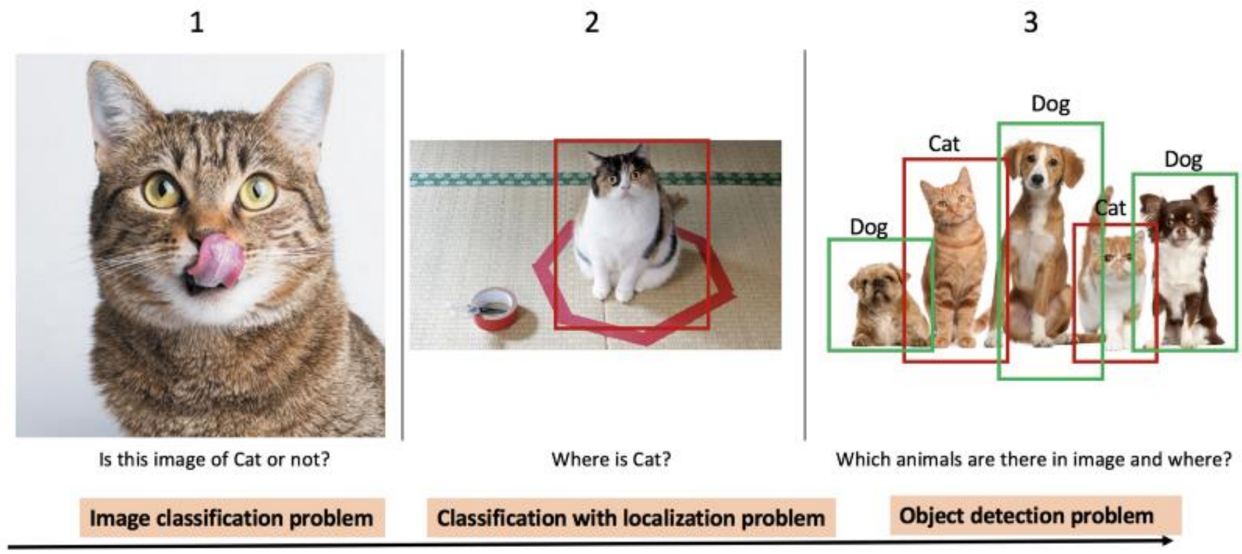


Figure 2.8. Popular computer vision tasks

The image shown in Figure 2.8, represent the most popular tasks carried out by computer vision modeling algorithms:

1. Image classification: The process of classifying images on the basis of recognized features or characteristics is known as classification. The process normally involves recognition of the dominant content in a scene. The dominant content gets the strongest confidence score irrespective of the transformation of that content such as scaling, location or rotation as can be seen in Figure 2.9.

For example, if we take the image of a cat or anything else, we want to know what dominant content is there. Thus, a classification system should always label that image as "cat" no matter where the cat is in the image so long as the cat is the dominant content there. The system should change the label of the image to the next dominant content, if the cat is no longer the dominant content.

.

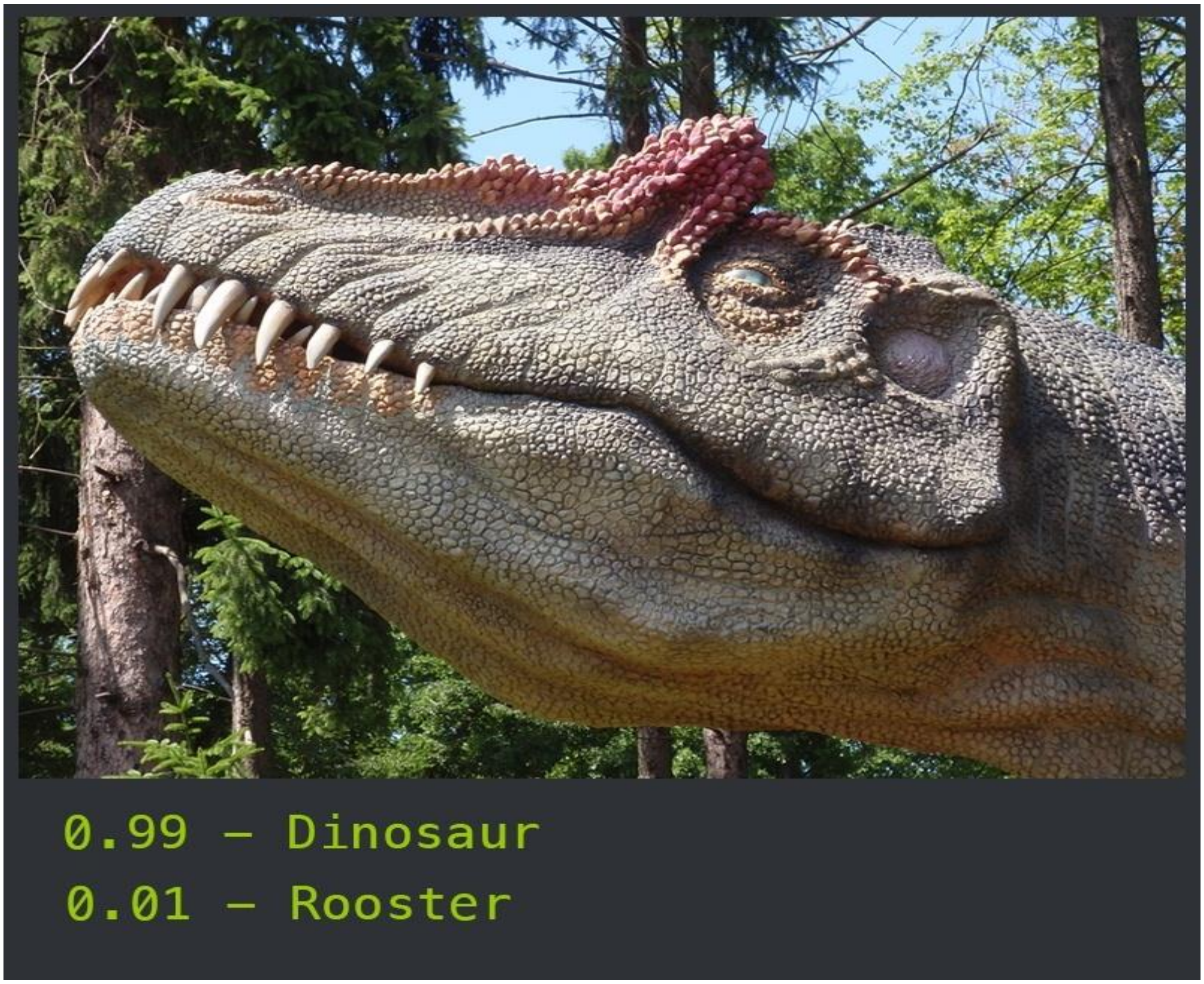


Figure 2.9. Classification example

2.Object classification and localization: Now sometimes we need to know how many dogs or how many known objects (objects in the database) are there and where they are in a scene. The process of recovering the location of the objects is called localization this can be shown in Figure 2.10.

Localization and classification are similar to each other, the only difference is the localization process finds the location of a single object of interest inside the image. (Javier, 2017).

Figure 2.10. Localization process example

For different useful real-life problems localization can be used. Such as, smart cropping (locating where the object is and crop the images according to object location), or the technic used for regular object extraction. Combining object localization with classification for not only finding the location of the object but also categorizing it into one of many possible categories.

3. Object detection: For classifying multiple objects at the same time, we will need to combine localization and classification together, this combination will end up with the need for detecting. Object detection is stand for finding and classifying a different number of objects in the image Figure 2.11 is an example for object detection.

Thus, object detection may be represented as:

detection=classification + localization.

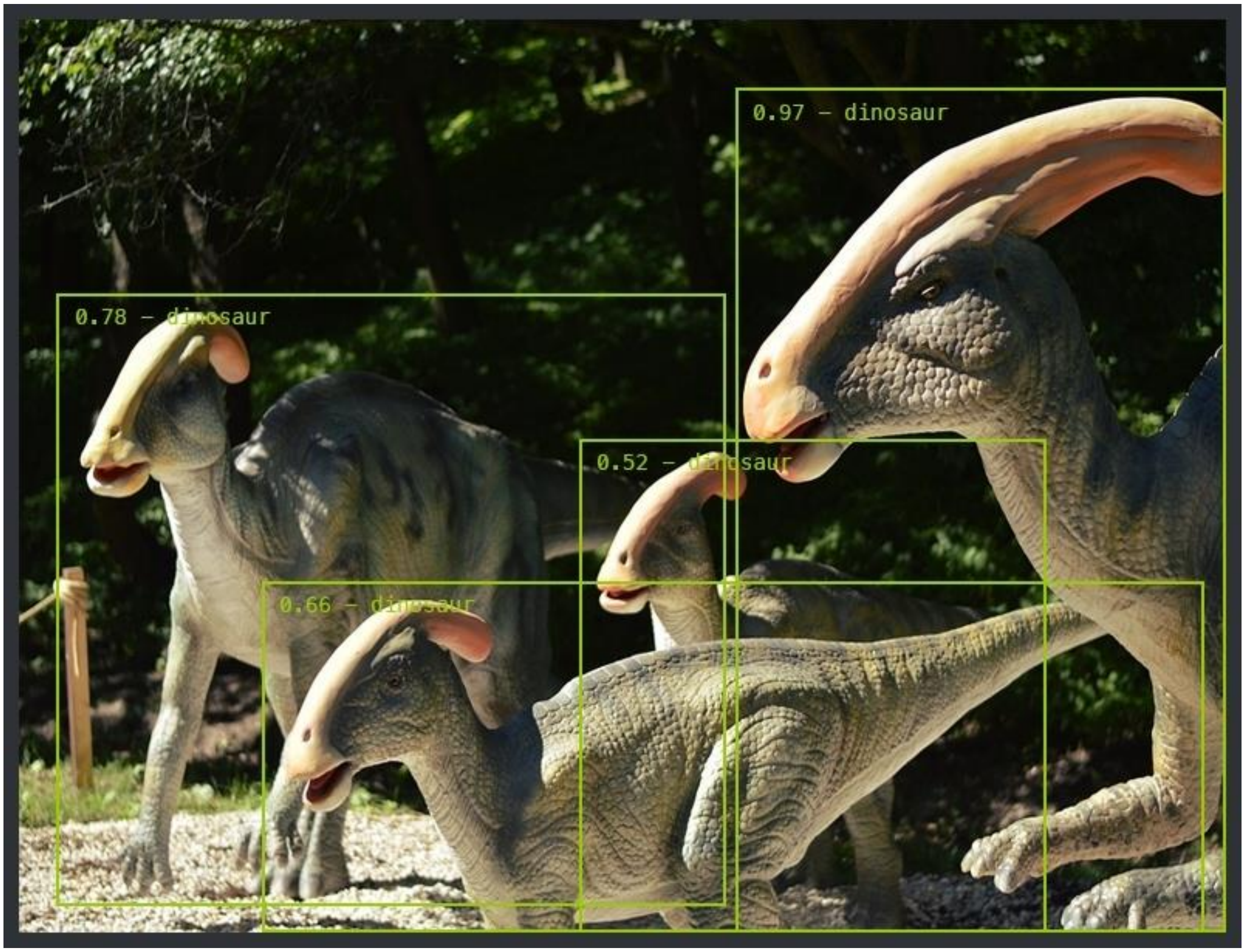


Figure 2.11. Object detection example

3. OBJECT DETECTION

Object detection is a main task in different computer vision. The main purpose of object detection is to classify multi object, and determine the placement of these objects in the input images or videos. The position of the object is specified by the bounding rectangle of the object or by a tight mask that preform the pixels corresponding to the objects.(Robert, 2017) .The difficult task in object detection problem is handle multiple probable appearances, deny irrelevant information which represented by the image noise and holding depressed execution time.

Object detection algorithm able to known group of objects might be present and locating their position inside the image for given image or a video stream.

Figure 3.1 shows how object detection model can identify two objects, the model has been recognized the two object and provide information about their positions:

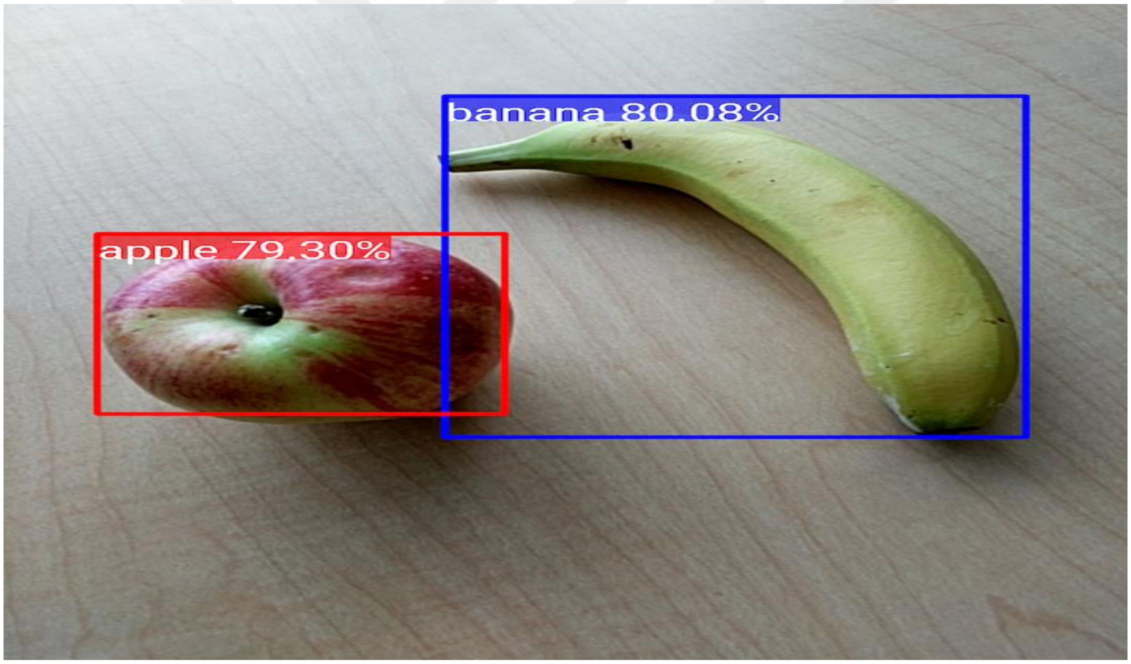


Figure 3.1. Finding object inside the image

The object detection algorithm can be used for detecting the existence and the position of various classes of objects. The model can be trained with images that include different segment of vegetables, over the label that assign the class of vegetables they perform (for example, a potato, a tomato, or a carrots), and data specifying where every object obvious in the image.

After the providing the image for the model, the output will be a list of the objects that the model detected, the bounding box location that hold each of these objects, and the confidence score that point that detection was correct.

3.1. Datasets for Object Detection

The **PASCAL Visual Object Classification (PASCAL VOC)** dataset consider one of the familiar datasets for object classification, object detection and objects segmentation. The PASCAL VOC contain approximately 10 000 images which used for validation and training, these images contain bounding boxes with the objects. although, it contains only 20 categories The PASCAL VOC dataset count as the object detection reference dataset. (Arthur, 2018).

The Microsoft developed **Common Objects in Context (COCO)** dataset which is detailed by Lin et al., (2014) by using per-instance segmentations the objects are labeled to aid in understanding an object's precise 2D location. Most of the algorithm concentrate within COCO object detection defy consisting in localizing the objects inside the image with bounding boxes and categorizing, because the dataset includes photos of 91 different objects gender that could be readily recognized along with per-instance segmentation masks. With a total labeled of 2.5 million instances in 328k images, Figure 3.2 contain different kind of images for several object.

Figure 3.2. COCO dataset (2015), segmented objects Examples

3.2. Performance Metric

The mean object detection tasks are finding all object in the image and drawing a **bounding box** can be define as a rectangle drawn around the object inside the image which tightly fits that object. For finding the exact boundaries of the objects of interest a process called **instance segmentation** can be used as shown in Figure 3.3**.**

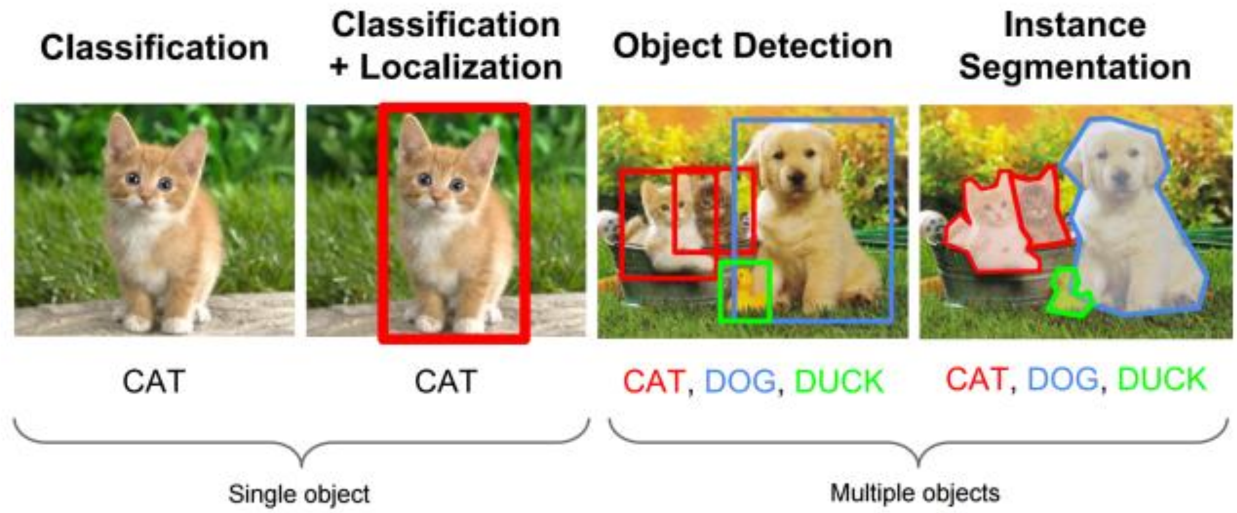


Figure 3.3. Different computer algorithm

The bounding box specifying object location. The bounding exists for every object in the image, it has 4 basic parameters (center of x, center of y, height, width). This can be trained by using a distance measure between predicted boxes and the bounding box of ground truth. (Kalshetti, et al., 2016).

According to Figure 3.4 The bounding box may be described by using different parameter:

1. Center of a bounding box (**b_x**, **by**)
2. The box width (**b_w**)
3. The box height (**b_h**)
4. The value of **(C)** is corresponding to a class of an object (car, traffic lights,).

Figure 3.4. Bounding box

In Figure 3.4. the p_c value represents a probability that there is an object in the bounding box. The object detection is a regression and a classification at the same time, the outputs model contains boxes more than the existing objects, in order to draw the bonding box around our object of interest low confidence boxes should be removed. This can be done by using **Intersection Over Union (IOU)** which represent evaluation metric used, it can be used by all the algorithm that's need to provides predicted bounding boxes in the output. The IOU value between 0 and 1 where the higher provide better predicted location of the bounding box for the object of interest. (Rosebrock, 2018).

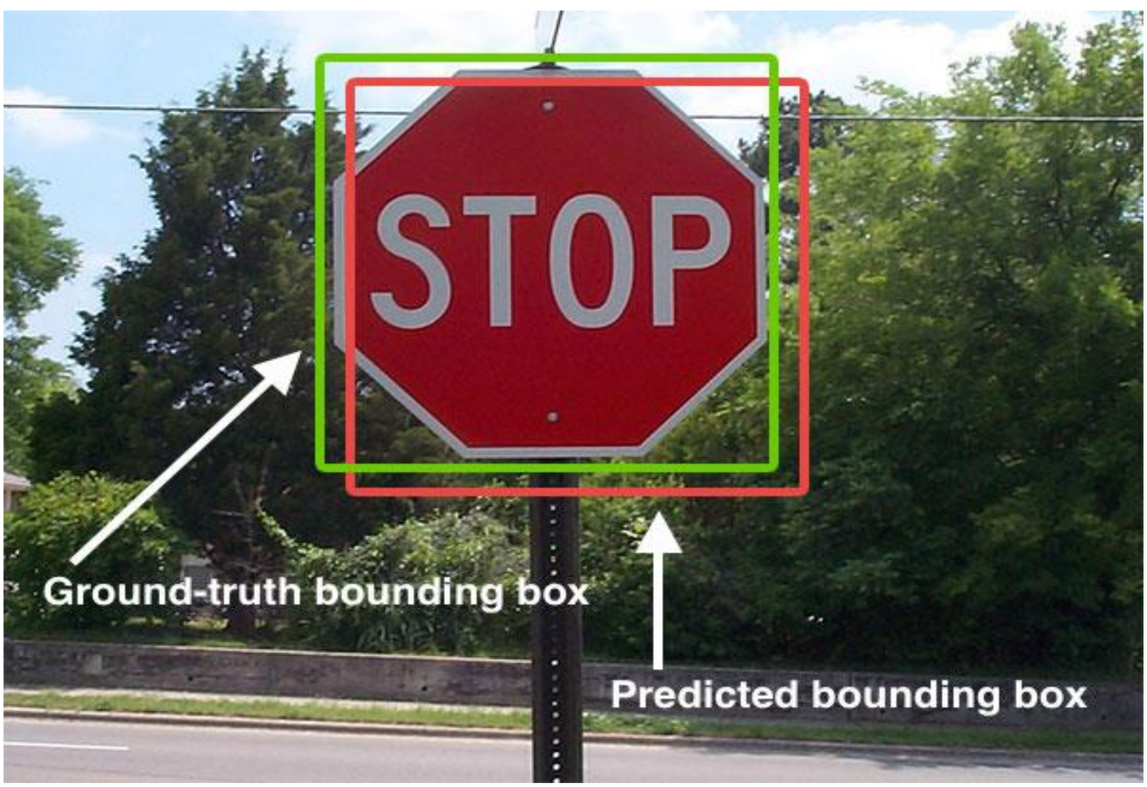


Figure 3.5. Example of IOU for detecting a stop sign in an image

In Figure 3.5. Show that the green box performs the bounding box for ground-truth, while the red box is symbolized predicted bounding box.

The Intersection over Union can be founded by dividing the amount of the area of overlap of the bounding boxes and the amount area of union:

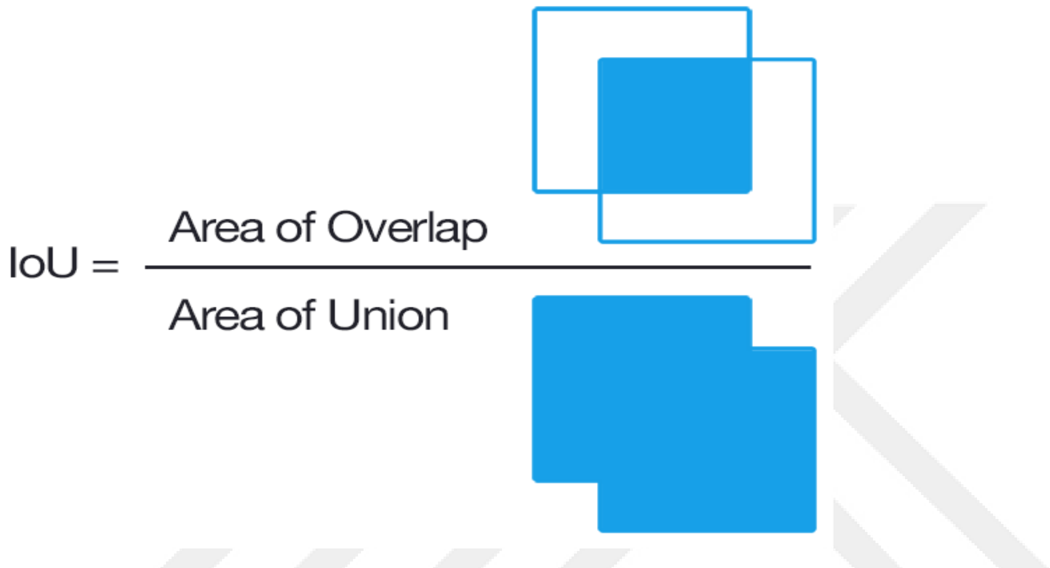


Figure 3.6. Intersection of Union representation

In Figure 3.6 we can see that the numerator represents the **area of overlap** which is between two values of the bounding box that the algorithm predicted and the bounding box for ground-truth.

While The denominator represents the **area of union**, the area is included the two boxes.

The Average Precision (AP) metric in binary classification the summary of the precision-recall curve. The **Average Precision (mAP)** is usually used metric for object detection which called the **mean**. The mAP represents the mean of the Average Precisions calculated through the challenge classes. The mAP metric averts to hold weak performances and maximum specialization in the classes. The mAP is mainly computed for a fixed IOU.

3.3. Non-Maximum Suppression (NMS)

Non-maximum suppression defined as a technique utilized for eliminating boxes that overlap so as to reduce the number of candidate boxes by an amount greater than a threhold. First sorting the boxes by using several criteria. Then going through the list of boxes that contain a number of boxes where the IoU overlap with the box under regard exceeds a threshold. Figures (3.7 and 3.8) show the difference between two approaches. In Figure 3.7 the result of applying NMS into the input image. In this Figure we use standard NMS (where the boxes are grouped by y coordinate of the bottom right corner). From Figure 3.7 shows that the boxes with minimum score being retained. In Figure 3.8 we use the modified NMS (where the boxes are grouped by foreground scores). While Figure 3.7 showing retained the boxes that have the highest foreground score, which is more eligible. In both cases, the overlap between the boxes will be assumed higher than the NMS overlap threhold.

Figure 3.7. The result of applying Standard NMS. Method

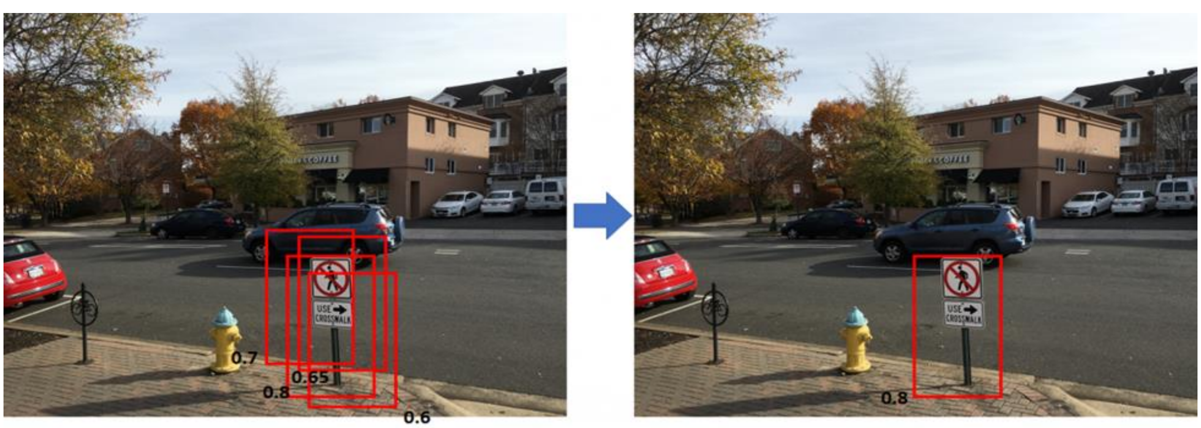


Figure 3.8. The result of applying modified NMS. Method

3.4. Object Detection Algorithms

3.4.1. Region-based convolutional network (R-CNN)

Selective search this method is improved by Uijlings and Arbelaez (2012**)**. This method tries to find object location by using an alternative to exhaustive search in the image. Selective search use hierarchical grouping to initializes small area in the image and combine them so as all object scales have to be taken into account. So as, the final combination is a box holding the whole image as shown in Figure 3.9. The areas that have been detected are combined according to a similarity metrics and set of color spaces. The output is a small number of proposals of the region that may be have an object by combine small regions.

Figure 3.9. The output image after applying Selective Search application

The R-CNN model developed by Girshick, (2015), Instead of trying to classify a huge number of regions, The R-CNN method merge the **selective search** method and the deep learning the selective search will be used for detecting region proposals, while the deep learning to discover the object inside these regions. To match the input of a CNN each region proposal will be resized. After resizing, a 4096-dimension vector of features will be extracted. For producing the probabilities that belong to each class, then, the features vector will be feed into multiple classifiers. There is an SVM classifier for each one of the classes, this SVM has trained to estimate a probability for detecting the object for a particular vector of features. This vector will be connected to a linear regressor to modify the bounding box shapes for a region proposal and this process will decrease the errors of localization process.

The steps that RCNN model follow it to detect the objects is listed below:

1. first step is taking the pre-trained convolutional neural network.
2. after retraining the model. According on the number of classes that need to be detected, the last layer of the network will be trained.
3. The third step is to get the Region of Interest for each image. The next step is resizing all regions so that they will agree the CNN input size.
4. After resizing the regions, we will train the SVM so as to classify objects and background. We will train one binary SVM, for each class,
5. The Final step is training the linear regression model, so as for each identified object in the image we will generate tighter bounding boxes.

- First, taken the image as an input the image shown in Figure 3.10 will be used as input image:

Figure 3.10. Input image

- By using some proposal method such as selective search, to get the Regions of Interest (ROI) this region shown in Figure 3.11:

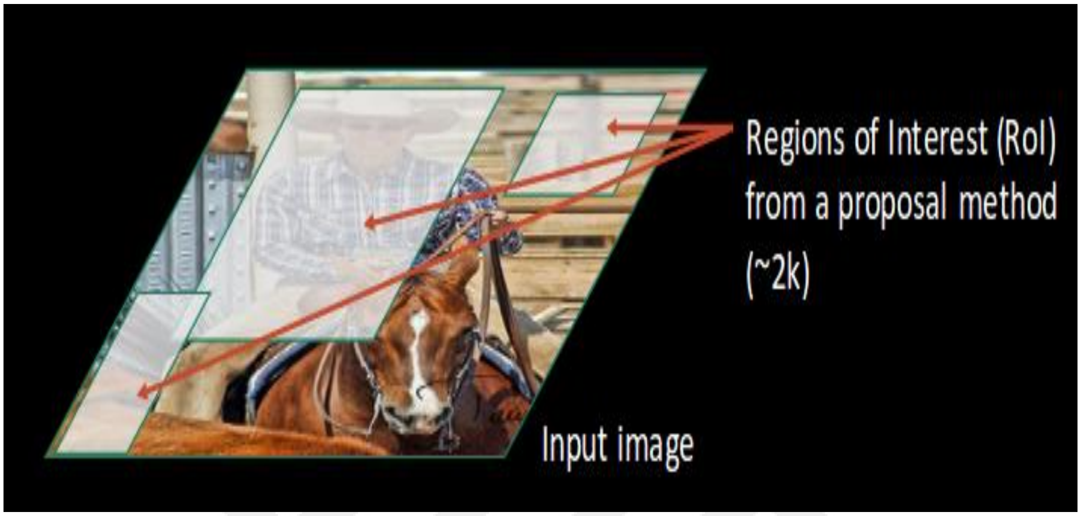


Figure 3.11. Regions of Interest (ROI)

- After getting the (ROI) then these regions will be reshaped to fit the input of the CNN, after that each region is passed to the ConvNet as shown in Figure 3.12:

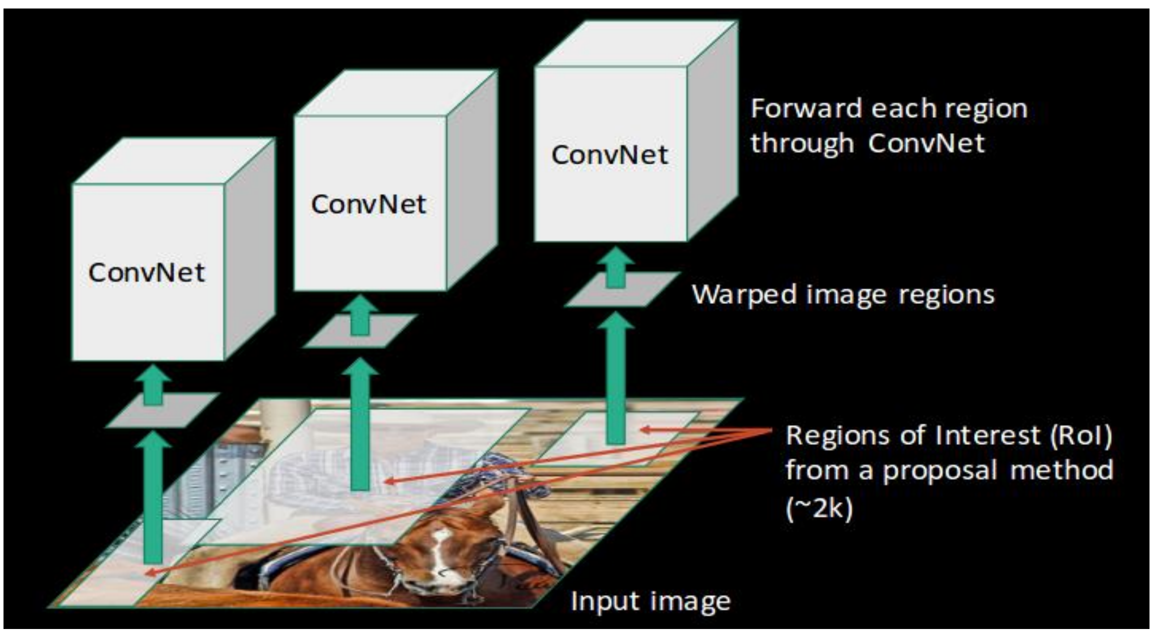


Figure 3.12. Passing the Regions of Interest (ROI) into the ConvNet

- Figure 3.13 shows how the CNN then extracts features for each region and then this region will be divided into different classes by using SVMs:

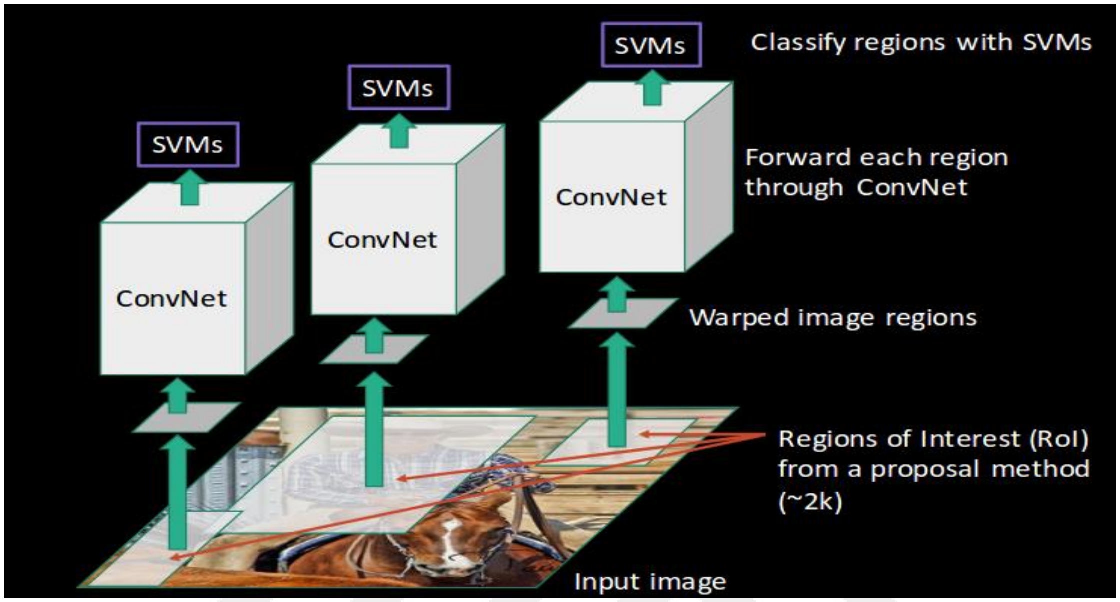


Figure 3.13. The regions are divided into different classes

The final step can be seen in Figure 3.14 the bounding boxes will be predicted for each identified region; this can be done by using bounding box regression (Bbox reg):

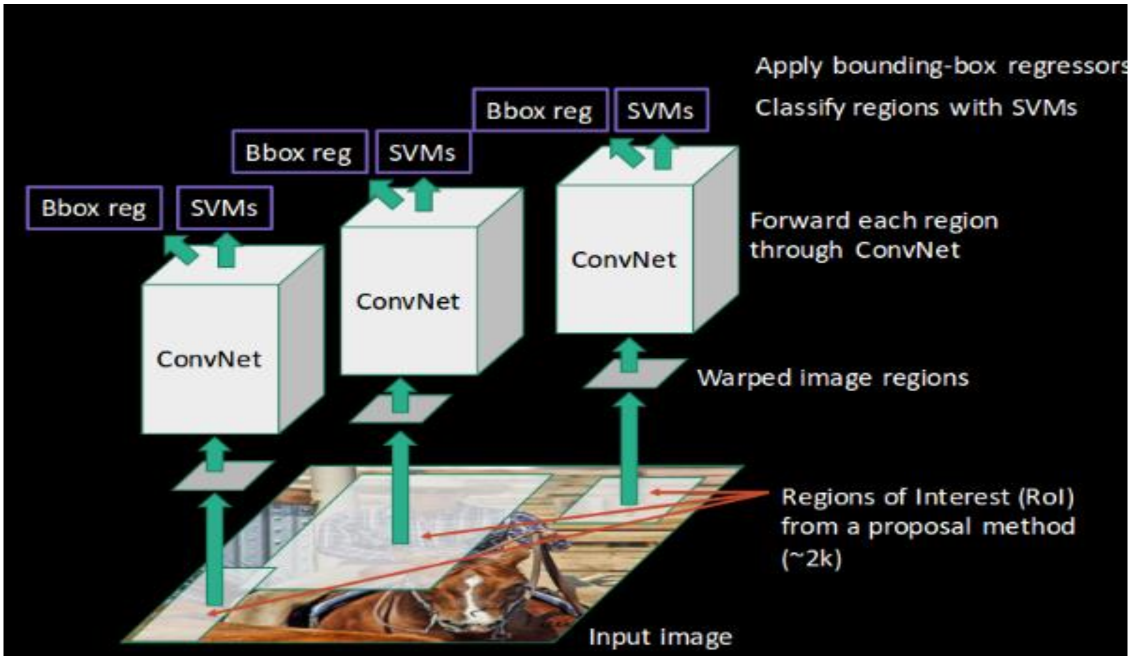


Figure 3.14. Predict the bounding boxes for each identified region.

3.4.1.a. RCNN problems

RCNN model used for object detection. But in order to train this model there is a number of limitations. One of these limitations is training the RCNN model is expensive and slow also:

- For using the selective search method for each image, we need to extract 2,000 regions.
- If we have N number of images, and we want to use CNN for extracting features in every image region. then the number of CNN features that we need it will be N*2,000.

All these processes work together and make RCNN very slow. for making the predictions every new image, the RCNN takes around 40-50. (Sharma,2018).

3.4.2. Fast R-CNN: fast region-based convolutional network

To decrease the consumption time because of the rise number of the models that required to analyses all the region proposals, Girshick (2015) developed the Fast Region-based Convolutional Network (Fast R-CNN).

instead of using a CNN for every region proposal like the (R-CNN) model. the entire image is taking as input by the main CNN with multiple convolutional layers. By applying the selective search method on the produced feature maps, we can detect the Region of Interests (RoIs). The RoI pooling layer will be used to obtain useful region of interests, these layers have a fixed height and width as hyperparameters, then the feature maps size is reduced. Creating a features vector need to feeds fully-connected layers for each RoI layer. The vector will use a classifier called SoftMax to predict the observed object, after that a linear regressor to adapt bounding box localizations these several processes shown in Figure 3.15.

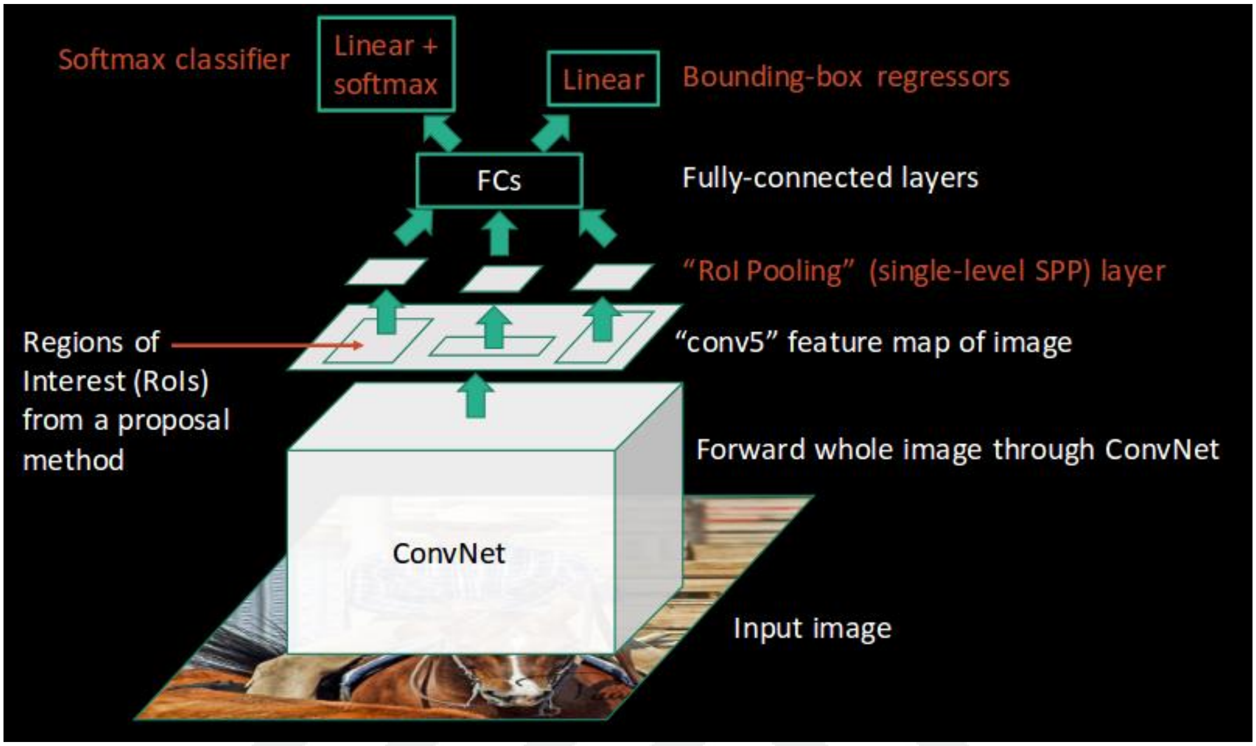


Figure 3.15. Fast R-CNN process

3.4.2.a. Fast RCNN. problems

To find the Regions of Interest, the fast RCNN uses selective search as a proposal method, which consider a slow and time-consuming process. For detecting objects, in the image It takes around 2 seconds, which is consider much better if we compared to RCNN.

3.4.3. Faster R-CNN: faster region based convolutional network

Fast RCNN uses selective search for generating Regions of Interest, which make the fast R-CNN slow and time consuming that affect the network performance. Shaoqing et al. (2016) developed the **Region Proposal Network** (RPN), this method eliminates the selective search algorithm, and directly generates region proposals, this lead to predict the bounding boxes and finding the objects faster. The Faster R-CNN combined between the Fast R-CNN model and the Region Proposal Network (RPN).

Similar to Fast RCNN the input of CNN model will carry on the entire image so as to produces the feature maps. Then the window that have a size of 3x3 slides will slides over whole feature maps so as to produce the features vector. This vector is related to two fully-connected layers, where one of these layers for box-regression and the other layer is for box-classification.

The output layer of the box-regression is fixed, it has a (4K)size, this means that the coordinates of the boxes, the (height and width) are also fixed with 4k. on the other hand, the output layer of box-classification is also fixed with 2K. when sliding window detect the k region then it's called anchors.

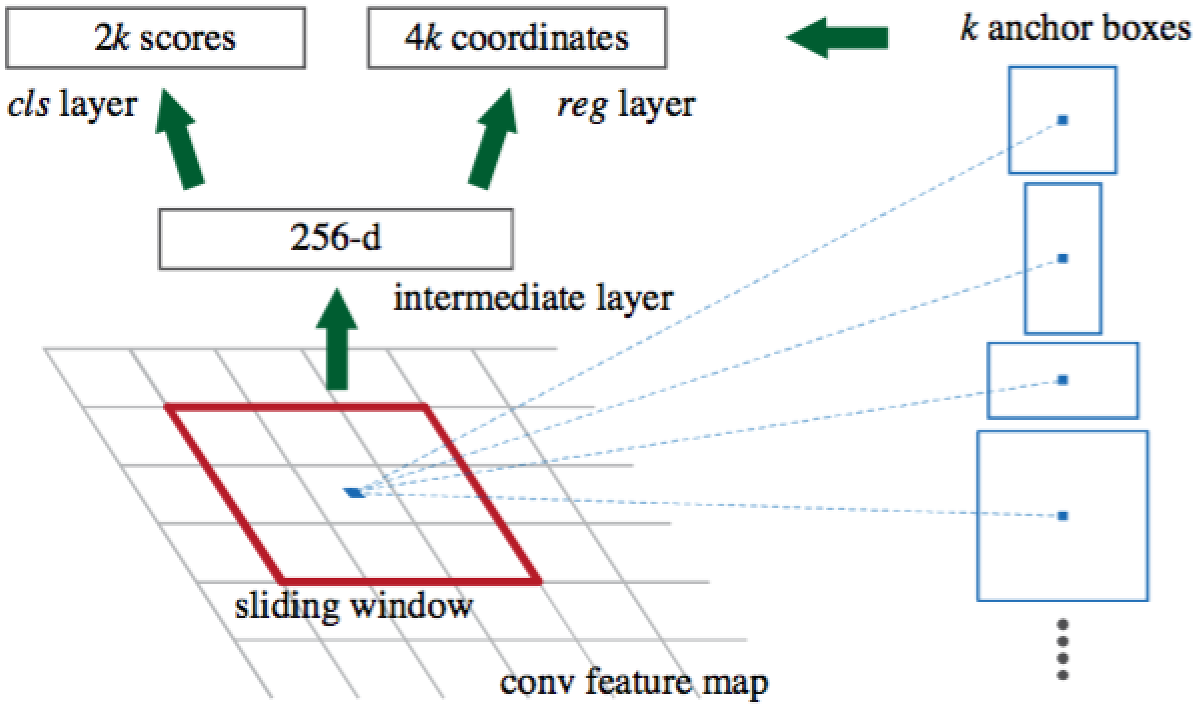


Figure 3.16. Single sliding window 3x3 size detecting the anchor boxes

Figure 3.16 show how to use the initial CNN model and feed it to Fast R-CNN model, with a view to compute the feature maps and the anchor boxes.

for improving the performances and accelerate the training, the faster R-CNN uses RPN so as to avert using selective search method this can be seen in Figure 3.17.

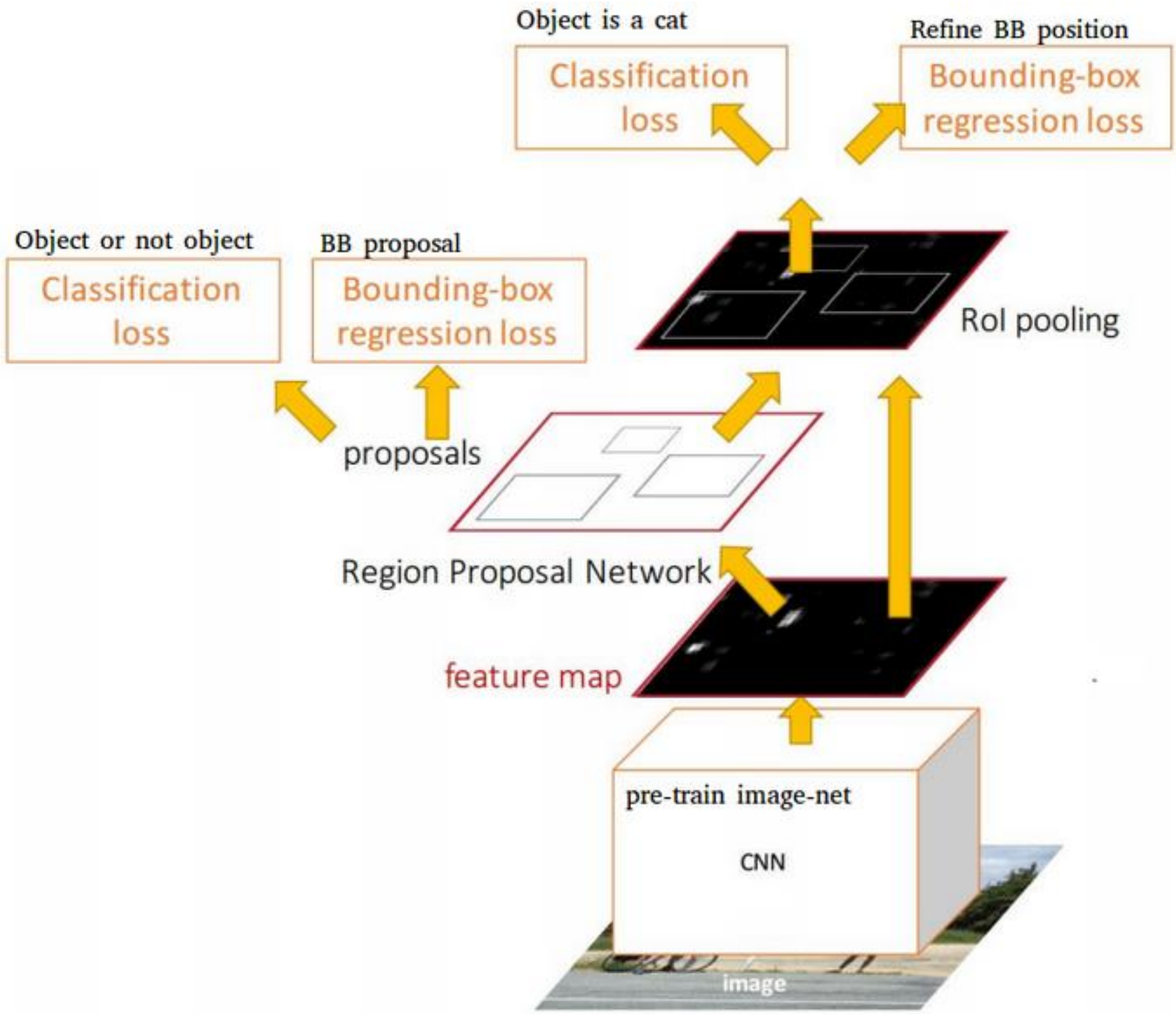


Figure 3.17. Fast R-CNN entire process

3.4.3.a. Faster RCNN problems

- to extract all the objects the in a single image, the algorithm requires too many passes.
- the interpretation of the systems builds on how the preceding systems performed because there are more than one system working in accordance with the other.

3.4.4. Summary of the algorithms covered

Table 3.1 contain the sum of the concept of object detection algorithms covered above.

Table 3. 1. comparison between different algorithms of object detection:

Algorithm	**Features**	**Prediction time / image**	**Limitations**
CNN	The input image will be divided into various regions after that every region classifies to the different classes.	-	For predicting the object accurately, the algorithm will take a lot of regions this will increase the computation time**.**
RCNN	Also Use selective search method to produce regions. For each image extracts about 2000 regions.	40-50 seconds	Using the selective search method increase the computation time because every region will be push through the CNN separately.
Fast RCNN	the image will be pushed one time to the CNN so as to extracted feature maps. Selective search will be used for creating the predictions.	2 seconds	The computation time is high because of using Selective search.
Faster RCNN	Eliminate the selective search method and use region proposal network.	0.2 seconds	There are different systems working in accordance with the other, so as the object proposal takes time the interpretation of systems will rely on the previous step.

3.4.5. YOLO: you only look once

YOLO model developed by Joseph (2016), this model uses a monocular network in an elementary evaluation so as the algorithm will immediately will predicts the bounding boxes and probabilities of the classes. The simple step that used in the YOLO model allows for predictions in real-time.

The model as shown in Figure 3.18 divides the input image to the (SxS) grid. Every cell in this grid will predicts a number of bounding boxes and a specific confidence score. This confidence performs the probability to detect the object multiply by the Intersection over Union (IoU) between the bounding box that predicted and the bounding boxes for the ground truth.

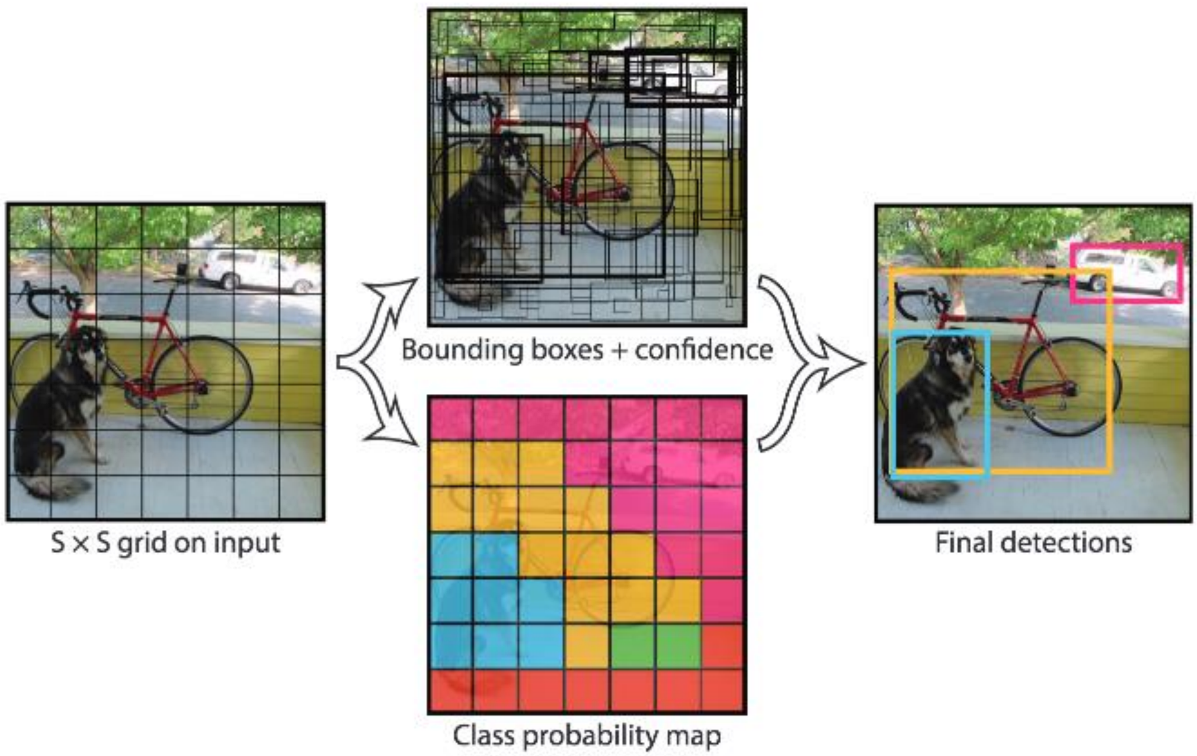


Figure 3.18. Application Example of YOLO object detection.

As we can see in Figure 3.19 the YOLO network contains 24 convolutional layers, these layers followed by 2 fully-connected layers. And then followed by reduction layers and (1x1) filters, these filters tread by convolutional layers (3x3), which will exchange the modules of premier inception.

The outputs layer the final layer contain S*S*((B*5) + C) tensor according to the predictions for every cell of the grid. Where B represent the anchor boxes number in each cell which is fixed, every one of these boxes has 4 coordinates (width and height, the center coordinates of the box) and a confidence value. C is the number of the probabilities estimated for each class.

The YOLO model able to predicts a very rise number of bounding boxes. Therefore, the output image contains many bounding boxes without any object. To eliminate this non necessary boxes YOLO model will use the **Non-Maximum Suppression** (NMS) method will be applied at the end of the network.

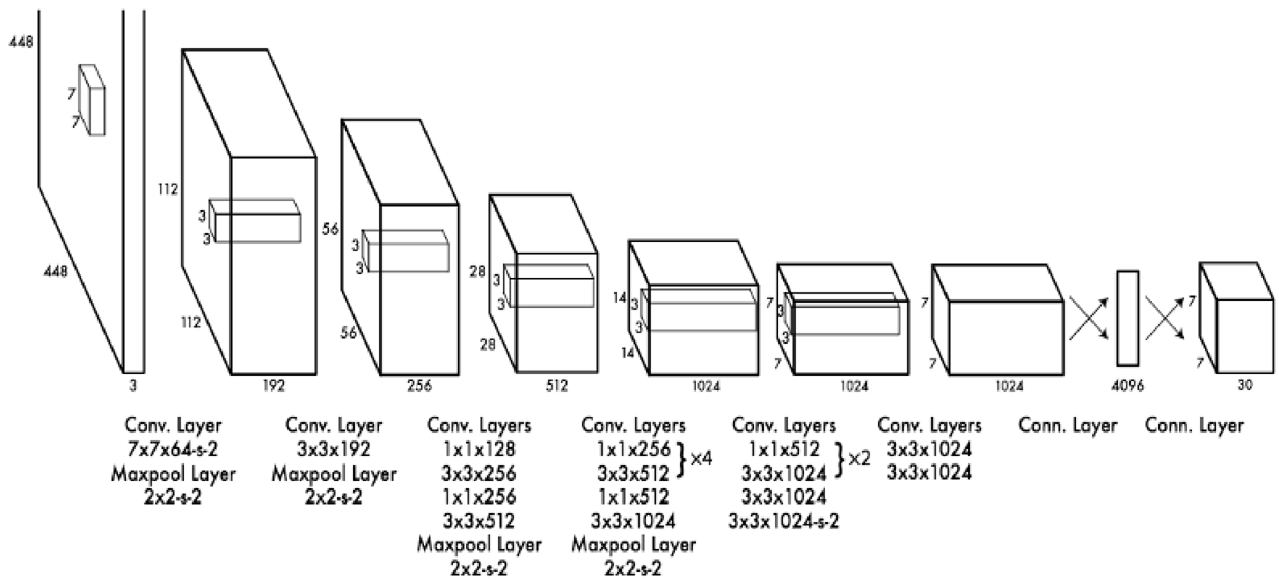


Figure 3.19. YOLO architecture diagram

In order to understand the difference between different types of YOLO algorithms Table 3.2 contain a comparison between these types.

Table 3. 2. Comparison between speeds and implementation for different YOLO algorithms trained with the datasets PASCAL VOC (2007) and (2012).

Model	mAP	FPS	Real Time speed
Fast YOLO	52.7%	**155**	Yes
YOLO	**63.4%**	45	Yes
YOLO VGG-16	66.4%	21	No
Fast R-CNN	70.0%	0.5	No
Faster R-CNN VGG-16	73.2%	7	No
Faster R-CNN ZF	62.1%	18	No

3.4.5. Limitations of YOLO model

YOLO has a constraint on bounding box predictions because every grid cell will have just one class and only predicts two boxes. This constraint limits the model ability to predict the nearby objects.

3.4.6. SSD: single shot detector

Liu and Anguelov (2016) improve a Single-Shot Detector (SSD) method, this method like the YOLO model it will predict whole at one time the bounding boxes and the probabilities of each class, the difference is that the SSD method will use an end-to-end CNN architecture.

The input image will pass over multiple convolutional layers, these layers has various sizes of filter (10x10, 5x5 and 3x3) as shown in Figure 3.20. For predicting the bounding boxes, the model will use feature maps from convolutional layers at several placement of the network. This bounding box are processed by a specific convolutional layer called extra feature layers which has 3x3 filters to product a group of bounding boxes identical to the boxes anchor of the Fast R-CNN.

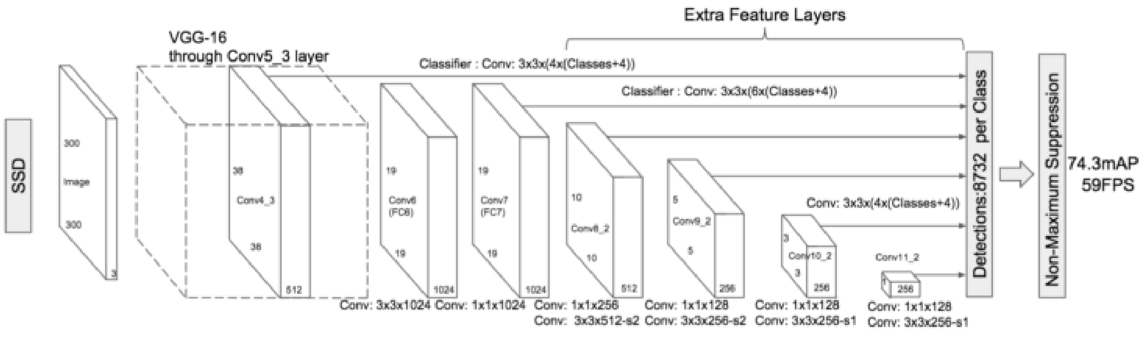


Figure 3.20. The SSD network model

Every bounding box contain 4 parameters: the width and the height, and the coordinates of the center. Figure 3.21 is an example for this coordinate.

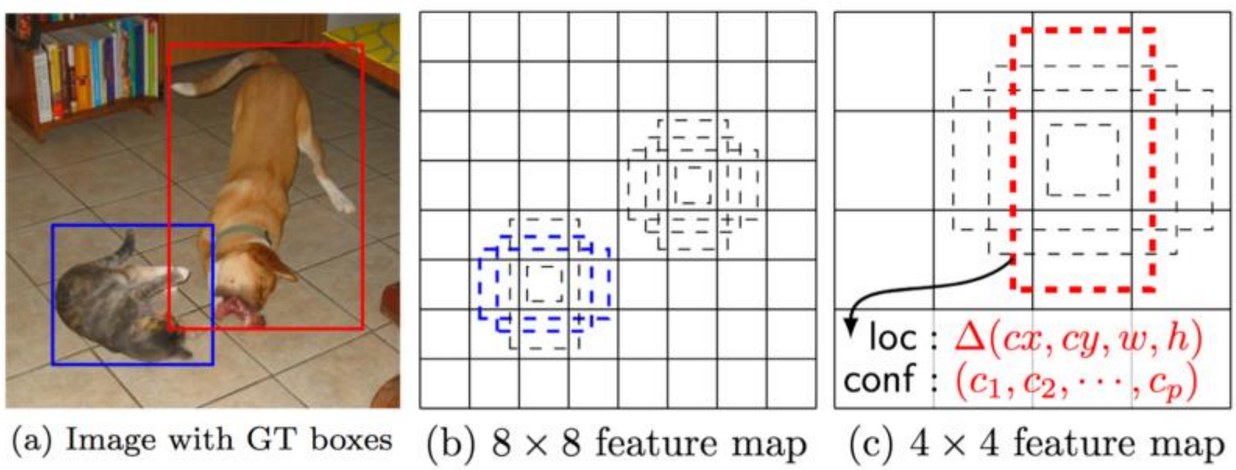


Figure 3.21. SSD Framework.

3.5. Practical Uses of Object Detection

There are various applications where object detection is demand. Detecting cancerous cells in medical applications which have great importance. Modern cars contain detection of traffic sign, that help in prohibit accidental lane changes.

3.5.1. Face detection

Face detection is a computer technology, this technic applied for different applications that require human faces identification in video or digital images. The technology may be applied in the tasks that need to find where the object is located and the sizes of that objects inside the image that belong to a given class. The face detection technology has the ability to detect frontal or near-frontal faces in a photo, regardless of lighting conditions or orientation, or skin color.

3.5.2. Counting

One of the important object detection application is counting. The capability to count cars, people, and flowers, in the images or video is demand for various types of systems such as video surveillance devices.

3.5.3. Visual search engine

The Pinterest visual search engine, one of the object detection application. The engine use object detection for indexing several sections of the image. To search for a particular purse, for finding instances of portfolio similar to that one who want in a various context. By using this technic, the shopping from internet become easier as shown Figure 3.22 by using object detection how to locate the product like shoes and bags. (Rey,2017).

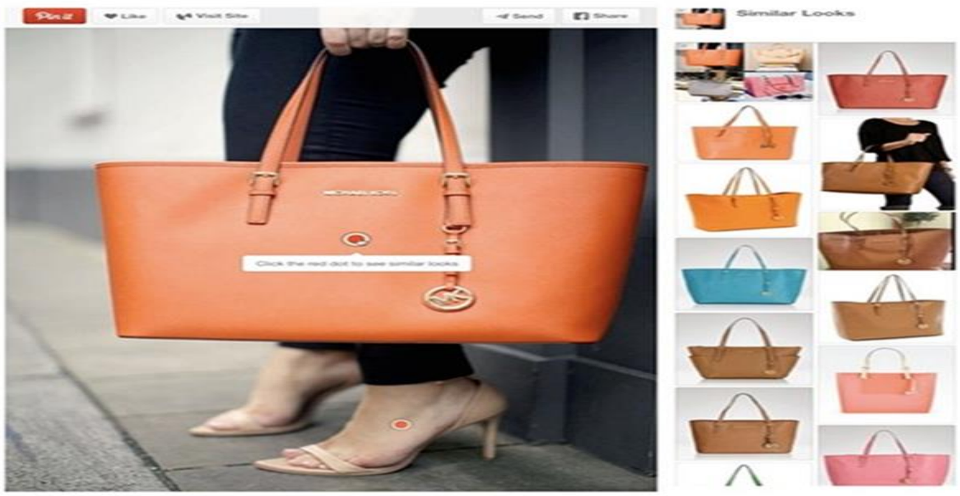


Figure 3.22. Applying object detection to localize products such as bags and shoes

3.5.4. Aerial image analysis

The inexpensive drones and satellite launch obtainable, made getting data from above is easy. Planet and Descartes Labs, apply object detection and satellite imagery to count ships, cars and trees. The result of this combination is great quality data, which was extremely expensive or impossible to obtain before.

Better View company use drone footage for automatic inspections to reach places difficult to reach it by human, Figure 3.23 is an example for this technology.

Figure 3.23. TensorFlight detection

4. EXPERIMENTAL RESULTS

The development of object classification and localization in image data, increased in recent years. to handle more and more complex image recognition problems, machine learning has become a major field of research. (Asplund, 2016). The safety of people is a subject of large interest to society. this thesis will try to find the best object detection algorithm, and apply this algorithm to the beach and swimming pools in order to guarantee people safety while they doing their activity on the beach or swimming pool.

The experimental work in this thesis is organized into two part in the first part we used YOLO algorithm for object detection and the second part we use SSD detection algorithm.

We will combine the object detection concept and object tracking into a single algorithm, to generate precise object detection algorithm, generally the algorithm includes:

Step1-Detecting:

1. detect the new objects that entered camera view
2. search for objects that maybe we lost them during the tracking phase.

Step 2 — Tracking: the algorithm has to phases, so as if it is not in the “detecting” phase, it will be in the “tracking” phase. The object tracker will track the object as it moves around inside the frame. In the algorithm the object tracker will operate more than the object detection algorithm, because the object detection algorithm much expensive, and take more consumption time, for that reason the object tracker should be faster and more efficient than the object detector.

the program, use both OpenCV and DLIB library. The OpenCV will be used for processing the input video. The object tracking algorithm will use centroid tracking. The first step for centroid tracker, is accepting the bounding box coordinates, after that the algorithm will use this coordinate to compute centroids as shown in Figure 4. 1. The bounding boxes coordinate may be provided by using YOLO or SSD detection algorithm.

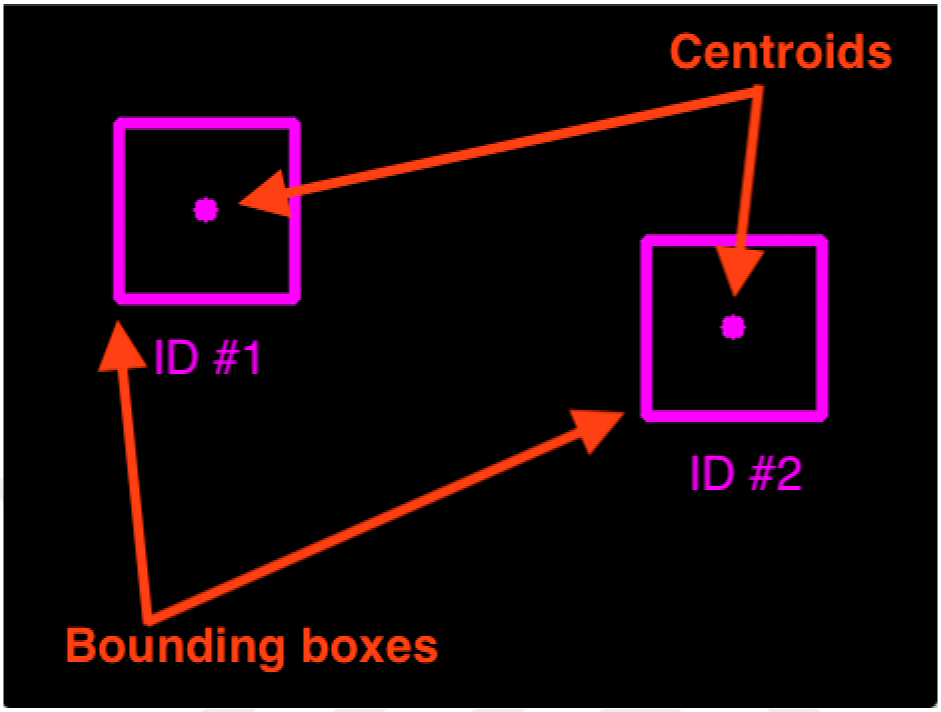


Figure 4.1. Bounding boxes and centroids

For object detection first we will use YOLOv3 and SSD object detection algorithms, YOLO object detector is one of the best real-time object detections, the algorithm can be obtained 45 **FPS** on a GPU. While, SSDs object detection algorithm originally developed by Google, SSD algorithm is also fast in the real-time object detection. It is also more straightforward than Faster R-CNNs.

Our project consists of three directories:

- object detector: The (YOLOv3 or SSD) object detector.
- videos: we'll process videos in real time.
- output: after processing input videos by (YOLO or SSD) and drawing bounding boxes around the object and returns the name of each class, the output can go into this folder.

The object tracker, will work then and accept the input *(x, y)*-coordinates of where the people is located in the video and will:

1. giving ID to that particular people. This ID will (swimmer or non-swimmer) corresponding to the people if he/she know how to swim or not.
2. Track that object whenever its moves around a video stream.

4.1. Algorithm

Input: to the input video file path.

1- import the necessary packages

2- build the argument parse:

--input= input video path

--output= output video path

--object detection algorithm= base path to (YOLO or SSD) directory

--threshold= threshold when applying non-maxima suppression

3-while loop on top of all the frames of the video stream file

4- For loop over each layer in the outputs

-- ensuring that the detection probability is more than the minimum probability by filter out weak predictions

-- according to the image size, the coordinates of the bounding box will be scaled

--finding the (x, y) the bounding box coordinates

-- the bounding box top and left corner of founded by using the coordinates of the center (x, y)

-- finding the centroid by using the bounding box coordinates to

cX=int((X+hight)/2.0)
cY=int((Y+width)/2.0)
input Centroids[i] = (cX, cY)

5- Drawing a special zone to classify people to swimmer and not a swimmer

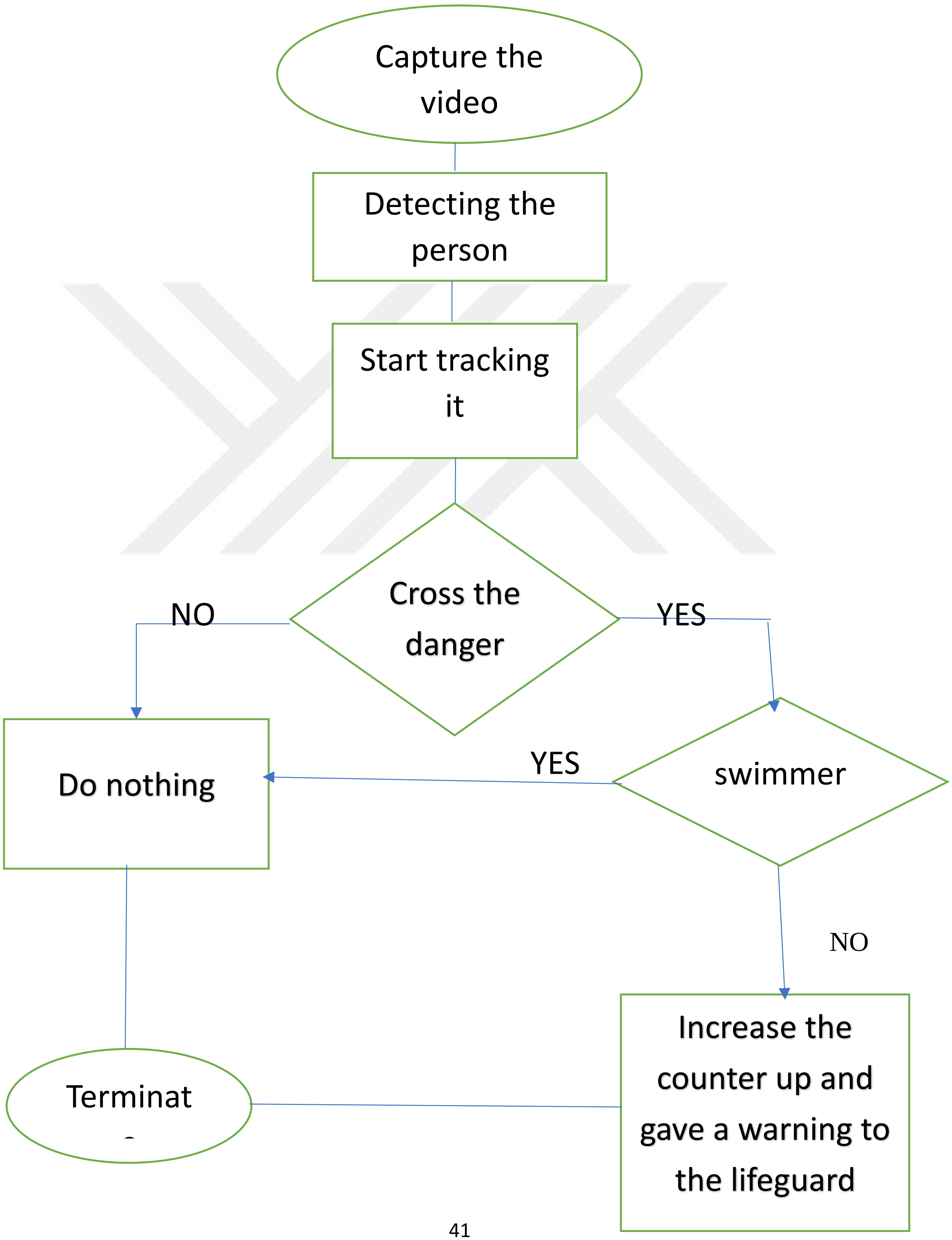
6- Drawing a line to sperate the safe from danger zone

7- For loop to check if someone cross the danger zone line

--If the person that cross the line is a swimmer continuo

--Else gave a warning to the lifeguards

4.2. Project Structure



4.3. Result

After applying YOLOv3 algorithm the experimental result is given in Figure 4.2, Figure 4.3 and Figure 4.4.

Figure 4.2. Experimental result with yolo algorithm test-1

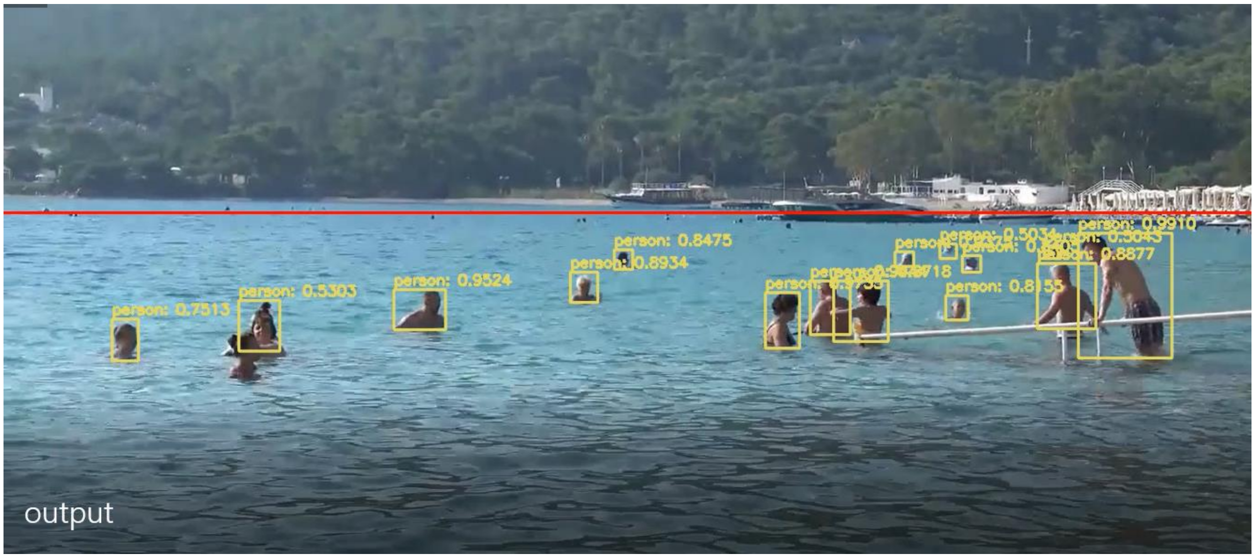


Figure 4.3. Experimental result with yolo algorithm test-2

Figure 4.4. Experimental result with yolo algorithm test-3

Figure 4.5 and Figure 4. 6 containing the experimental result after applying SSD detection algorithm.

Figure 4.5. Experimental result with SSD algorithm test-1

Figure 4.6. Experimental result with SSD algorithm test-2

After applying a different object detection algorithm, we reached a conclusion that the (YOLO and SSD) algorithms are the best a broach for our research. the most difficult part when all the human body is inside the water (deep water regions) only the head is out the water most of the algorithms cannot detect this person but when we apply the (YOLO and SSD) algorithms the result is very good in detecting these people and also the second problem is when the beach is too much corded.

According to figure 5.3 the red line represents the danger zone line if any detected person (person with yellow rectangle (not swimmer)) the program will give a warning for the lifeguard.

5. CONCLUSION AND RECOMMENDATIONS

The visual surveillance systems, represent the area of interest for human societies in different fields such as, public, commercial and research circles, military. (Hemangi and Bhagat, 2015). The mean goals of automated surveillance systems are to obtain detection, classification, recognition and tracking of the objects of interest. Applications of visual surveillance varies from the crime prevention, children at home, traffic control to monitoring patients at hospital. The field of intelligent surveillance are included in different domains such as computer vision, embedded systems, image processing, artificial intelligence.

Public safety represents the major concern for governments in the modern world. Visual surveillance and People detection, are key elements in the installations of event video monitoring, safety equipment is needed at public locations such as parks, streets, hospitals, etc.

This thesis focused on real-time object detection and object tracking. The proposed solution uses video cameras installed on land, which are capturing real-time videos of the sea. The use of video surveillance systems is popular practice whenever the safety is to be guaranteed. (Gomes, et al., 2014). Research display that lifesavers don't have the ability to watch everyone above and under the water for all the times. studies also show that lifeguards are completely involved in watching the water half the time.by using machine learning and video processing we billed computer algorithm, this algorithm use YOLO the fast, accurate object detector and centroid tracking algorithm to track the people after detecting them and put them inside a bounding box. This technology can be used in swimming poles and the beach to supplement lifeguards' job by providing 'eyes' watching the water at all the times – not some of the time. this algorithm technology has been developed to help lifeguards do their job to the best of their ability by alerting them quickly whenever the person crosses the line to the dangers zone.

there are three primary object detection methods:

- Faster R-CNNs developed by (Girshick, 2015)
- (YOLO): You Only Look Once developed by (Redmon, 2015)
- (SSDs): Single Shot Detectors developed by (Liu, 2015)

Faster R-CNNs, one of the technics that used for object detection, but this technique can be hard to implement, difficult to understand and challenging to train. Moreover, implementation the faster R-CNNs the algorithm also quite slow.

for increasing the speed of deep learning-based object detestation and tracking, we will use Single Shot Detectors (SSDs) and YOLO, the strategy that these two algorithms use it is detecting the object in one-stage. These algorithms deal with object detection like a regression issue, after taking the input image and learn the coordinates of bounding box and the probabilities for each class label with each other. Generally, using single-stage detectors algorithm will be less accurate than two-stage detectors but the one-stage are faster. (Rosebrock,2018).

The largest limitation and obstacle of the YOLO object detector is that:

1. YOLO algorithm does not detect small objects completely at all the time
2. YOLO algorithm does not detect objects that the distance between them is too much small

On the other hand, the other limitation is the deep learning object detection algorithm are very computationally expensive, especially when this algorithm run on the CPU. To overcome on this limitation, the program will use a combination of the object detection algorithm and object tracking into a single algorithm, the benefit of this hybrid approach is that it can be applied the best algorithm of object detection (YOLO or SSD) for a period so as to get the exact coordinate of the people, after that the object tracker will get these coordinate and use them to find the centroid for each person in the beach and then it will start tracking that person .

After applying the hybrid algorithm, the most difficult part is the Occlusion handling. Occlusion is perhaps the most critical challenge in MOT. It is a primary cause for ID switches (swimmer and non-swimmer) or fragmentation of trajectories. In order to handle occlusion, we will compare to detection position, the current position and the position for the next frame. That can be done by determining the relative distance, if the two objects are closed to each other so as to match them.

REFERENCES

Arthur, O., (2018). Review of Deep Learning Algorithms for Object Detection. Date Accessed: 5.02.2018. http://Medium.com.

Asplund, C., (2016). Object Classification and Localization using machine learning techniques. University of Technology Gothenburg, Department of Physics Chalmers, Master's thesis in Complex Adaptive Systems, 63p, Sweden.

Dantone, M., (2014). Real Time Detection of Human Body and Face Parts, Medieninformatiker, Ludwig-Maximilian University, PHD. Thesis,172p, Germany.

Gonzales, R., Woods, R., (2008). Digital Image Processing, Prentice Hall, 954, New Jersey.

Girshick, R., (2015). Fast R-CNN, IEEE International Conference on Computer Vision (ICCV), Pages 1-9.

Gomes, P., Santana, P., Barata, J., (2014). A Vision-based Approach to Fire Detection, International Journal of Advanced Robotic Systems, 11 (149), 1-2.

Hashemia, N., Aghdamb, R., Ghiasic, A., Fatemid, F., (2016). Template Matching Advances and Applications in Image Analysis American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS**)** Volume 26, No 3, pp91-108.

Hemangi, R., Bhagat, K., (2015). Detection and tracking of moving object: A Survey. Int. Journal of Engineering Research and Applications, 5(11):138–142.

Hu, W., Tan, T., Wang, L., Maybank, S., (2004). A survey on visual surveillance of object Motion and behaviors, in IEEE Transactions on Systems, Man and Cybernetics, vol. 34, pp. 334_352.

Javier, W., (2017). Object Detection with Deep Learning: The Definitive Guide. Date Accessed: 30.08.2017. http://tryolabs.com.

Jalled, F., (2016). Object Detection Using Image Processing, Moscow Institute of Physics & Technology, 07791(1611), 1-3.

Joseph, R., Ali, F., (2016). You Only Look Once: Unified, Real-Time Object Detection. Rapor No: 1506.02640v5.

Kawahito, S., Seo, W., (2017). Noise Reduction Effect of Multiple-Sampling-Based signal-Readout Circuits for Ultra-Low noise CMOS Image Sensor. Sensors, 16(11), 1-19.

Kalshetti, P., Jaiswal, A., Rastogi, N., Gangawane, P., (2016). Object Detection, 16p.

Kloss, A., (2015). Object Detection Using Deep Learning. Eberhard Karls University, Faculty of Mathematics and Natural Sciences, Master Thesis Computer Science, 101p, Germany.

Martin, A., (2009). People Detection Algorithms Based on Appearance and Motion Information, Department of Computer Engineering, Autonomous University of Madrid, Master Thesis, 88p, Spain.

Navneet, D., (2006). Finding People in Images and Videos, Mathematics Science and Information Technology, PHD. Thesis,135p, France.

Piccardi, M., (2004). Background subtraction techniques: a review, IEEE International Conference on System, 1-4.

Rajkamal, K., (2014). Object detection and tracking in video image. National Institute of Technology Rourkela. Master thesis. India.

Rey, J., (2017). Object Detection with Deep Learning: The Definitive Guide. Date Accessed: 30.08.2017. https://tryolabs.com

Robert, V., (2017). Object detection based on candidate generation and classification. Technical University of Cluj-Napoca, Computer science category, PHD. Thesis, Romanian.

Rosebrock, A., (2018). Tutorials in YOLO object detection with OpenCV. Date Accessed: 12.12.2018. https://pyimagesearch.com.

Sharma, P., (2018). A Step-by-Step Introduction to the Basic Object Detection Algorithms (Part 1). Date Accessed: 11.10.2018. https://analyticsvidhya.com

Shaoqing, R., Kaiming, H., Girshick, R., Jian, S., (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence,USA.

Sulaiman, S., Alias, M., Isa, N., Rahman, M., (2007). An Expert Image Processing System on Template Matching. International Journal of Computer Science and Network Security, VOL.7 No.7.

Uijlings, J.,Arbelaez, E., (2012). Selective Search for Object Recognition. Springer International Journal of Computer Vision, New York.

Valera, M., Velastin, S., (2005). Intelligent distributed surveillance systems: a review, in IEE Proceedings on Visual Image Signal Processing, vol. 152, pp. 192_204.

Liu. W., Anguelov, D., (2016). SSD: Single Shot MultiBox Detector. Proceedings of the European Conference on Computer Vision (ECCV), pages 1-5.

Lin, T., Deva, R., Maire, M., Belongie, S., Hays, J., Perona, P., Dollar, P., Zitnick, C., (2014). Microsoft COCO: Common Objects in Context, Springer International Publishing, LNCS 8693, pp. 740–755.

Yilmaz, A., Javed, O., Shah, M., (2006). Object tracking: A survey, ACM Computing Surveys, 38(4), 1-4.

CURRICULUM VITAE

Name and Surname	:	Zahraa Qosay Hashim ALHAIDARI
Place and Date of Birth	:	MOSUL, 02/07/1991
Marital Status	:	Married
Foreign Languages	:	English (Advanced)
		Turkish (Beginner)
E-mail	:	zahraahashem91@gmail.com

Educational Status

Bachelor's Degree	:	Mosul University Faculty of Engineering Mechatronics Engineering, 2013
Master's Degree	:	Istanbul Ticaret University Graduate School of Natural and Applied Sciences Department of Mechatronics Engineering, 2019

Publications

A. Ozpinar, Z. Alhaidari, "A Machine Learning Based Video Processing Warning System for Lifeguards", 3rd International Symposium on Innovative Approaches in Scientific Studies, 19-21 April 2019, Ankara