University of Kasdi Merbah Ouargla
Faculty of New Information Technologies
And Communication

A Thesis Presented to the Department of
Computer science and information technologies
For the Degree of Master LMD in
……

By**: Ghofrane BENDEBKA**

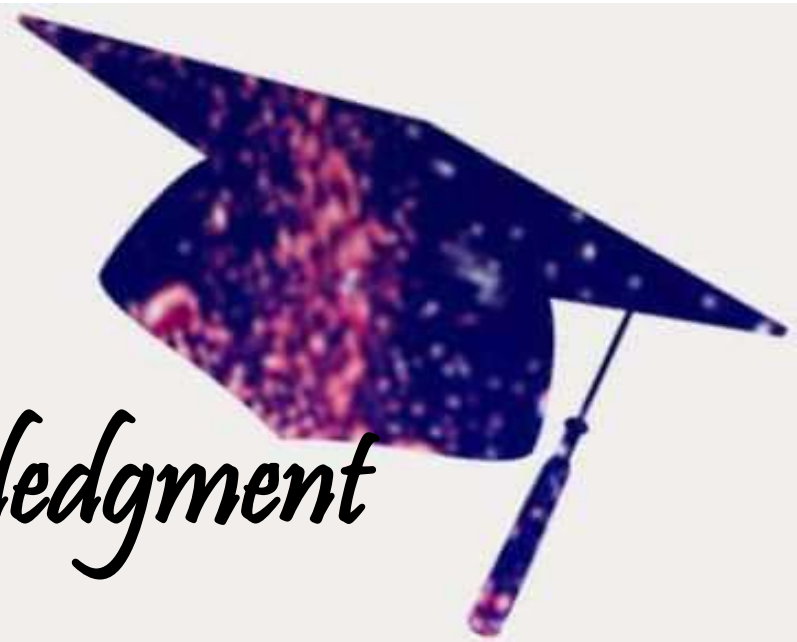
Title

Interpretation and semantic indexing of an image content

Members of jury:

Dr.	Adel ZGA	University of Ouargla	Supervisor
Dr.	Akram Boukhamla	University of Ouargla	Examiner
Dr	Djalila Belkebir	University of Ouargla	Examiner

OUARGLA: 01/10/2020



Acknowledgment

In the name of God and prayer and peace be upon the Messenger of Allah. Praise be to God, and bless us for where we are today. Thank God who blessed me with joy today after a long academic journey. Thanks to those who helped and supported me with all strength and patience to complete this work professionally

Thank you to the supervising professor Adel Zga for his efforts to follow up on my achievements in this valuable work

Also, I do not forget the merits of all my professors who contributed to my studies until the end of my academic career since its beginning Thank you all .

Thanks to everyone who taught me a letters one day

Dedication

I dedicate this project to my parents Abderrazak, Yamina who were the way to my success and my steadfastness, and I offer to dedicate my success and my work to them to return a little of what they did for me.

To my younger sister Nabila.

To My childhood companions who have been helping.

To My companions who have been a source of trust for me, as they were and still are my strength that will never let me down.

I also do not forget those whom God passed away with my sincere prayer, especially my aunt Bakhta who raised me and the all victims of the Corona.

Also to my dear cousin Roufida

Abstract

The field of object detection has received a lot of interest from researchers in recent decades. As it is estuary for many interesting applications, in many areas of life such as unmanned aerial vehicles, self-driving cars, robots, and many other applications that require recognition of objects and scenes within images.

Due to the large number of crimes and work accidents in the last century, our goal is to propose a method for introducing the concept of self-monitoring in the field of security, for example, for a group of people, each individual has his or her unique appearance and body shape, this is what makes it difficult to recognize, and we find that there is still no acceptable approach that gives an accurate semantic interpretation of images compared to the performance of humans, which made access to classifying the dangers to which a person may be exposed with high accuracy is difficult.

We created a new dataset dedicated to work accidents. We used YOLOv3 model to verify the presence of humans in the images, then we classified them into two categories (the category of presence of risk and the other being the absence of risk) using the CNN (vgg16 and inception v3) and KNN algorithms. In general, our methodology will provide additional information for the semantic analysis and interpretation of the content of the images, which is our desired result. The result Obtained it is acceptable and promising, so the proposed method helps save lives, but it needs more data.

Keywords: object detection, deep learning, HOG, SVM, YOLO,

Résumé

Le domaine de la détection d'objets a suscité beaucoup d'intérêt de la part des chercheurs au cours des dernières décennies. Comme il est un estuaire pour de nombreuses applications intéressantes, dans de nombreux domaines de la vie tels que les véhicules aériens sans pilote, les voitures autonomes, les robots et de autres applications qui nécessitent a la reconnaissance d'objets et de scènes dans les images.

En raison du grand nombre de délits et d'accidents du travail au dernier siècle , notre objectif est de proposer une méthode pour introduire le concept d'auto-surveillance dans le domaine de la sécurité, par exemple, pour un groupe de personnes, chaque individu a son apparence et sa forme corporelle uniques, c'est ce qui la rend difficile à reconnaître, et nous constatons qu'il n'y a pas d'approche acceptable qui donne une interprétation sémantique précise des images par rapport à la performance des humains, ce qui a permis de classer les dangers auxquels une personne peut être exposée avec une grande précision est difficile.

Nous avons créé un nouvel ensemble de données dédié aux accidents du travail. Nous avons utilisé le modèle YOLOv3 Pour vérifier la présence d'humains dans les images, nous les avons ensuite classées en deux catégories (la catégorie de présence de risque et l'autre étant l'absence de risque) en utilisant l'algorithme CNN (Vgg16 and Inception v3) and KNN. En général, notre méthodologie fournir des informations supplémentaires pour l'analyse sémantique et l'interprétation du contenu des images, ce qui est le résultat souhaité. Le résultat obtenu est acceptable et prometteur, la méthode proposée permet donc de sauver des vies, mais elle a besoin de plus de données.

Mots clés: détection d'objets, apprentissage en profondeur, HOG, SVM, YOLO

الملخص

حظي مجال اكتشاف الأشياء باهتمام كبير من طرف الباحثين في العقود الأخيرة . حيث إنه يعد مصبا للعديد من التطبيقات المثيرة للاهتمام، في العديد من المجالات الحياة مثل الطائرة بدون طيار، والسيارات ذاتية القيادة، الروبوتات ، والعديد من التطبيقات الأخرى التي تتطلب التعرف على الأشياء والمشاهد داخل الصور .

و نضرا لكثرة الجرائم وحوادث العمل في القرن الأخير هدفنا هو اقتراح أسلوب لإدخال مفهوم ذاتية المراقبة في مجال الامن سنحاول التركيز في هذه الأطروحة على إنشاء نظام فعال للتعرف على الانسان وتصنيفه ما اذا كان يتعرض لخطر التي تعتبر مهمة صعبة على سبيل المثال بالنسبة لمجموعة الأشخاص لكل فرد مظهره وشكل جسمه الفريد و هذا ما يجعلها من الصعب التعرف عليه كما نجد انه لا يوجد حتى اليوم نهج مقبول يعطي تفسير دلالي دقيق لصور مقارنة بأداء البشر مما جعل الوصول لتصنيف الاخطار التي قد يتعرض اليها انسان بدقة عالية امرا صعبا. قمنا بإنشاء قاعدة بيانات جديدة مخصصة لحوادث العمل استعملنا نموذج yolo v3 لتحقق من وجود انسان في الصور ثم قمنا بتصنيفها الى فئتين(فئة وجود خطر و الأخرى عدم وجود خطر) باستعمال خوارزمية (KNN , CNN (Vgg16 , Inception V3 .
بشكل عام, ستمنح منهاجيتنا معلومات اضافية لتحليل و التفسير الدلالي لمحتوى الصور و هي النتيجة المرجوة لدينا .وكانت النتائج التي تم الحصول عليها مقبولة وواعدة، وبالتالي فإن الطريقة المقترحة تساعد على انقاذ أرواح الناس لكنها تحتاج لمزيد من البيانات.

الكلمات المفتاحية: اكتشاف الأشياء , التعلم العميق , YOLO , SVM , HOG .

Contents

List of Figures

List of tables

List of Abbreviations

AP	Average Precision.
ANN	Artificial Neural Network
CNN	Convolutional Neural Network.
COCO	Common objects in Context.
CONVNET	Convolutional Network.
DCD	Dominat Dolor Descriptor
HOG	Histogram of Oriented Gradients
HTD	Homogeneous Texture Descriptors
KNN	Kearst Neighbor Network
ML	Machine Learning
PASCAL VOC	PASCAL Visual Object Classes.
R-CNN	Region-based Convolutional Neural Network
ResNet	Residual Neural
R-FCN	Region-based Fully Convolutional Network.
ROI	Region Of Interest.
RPN	Region Proposal Network.
RSD	Region-based Shape Descriptors
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine
SSD	Single-Shot Detector
TBD	Texture Browsing Descriptors
YOLO	YOU Only Look Once
3D SD	3-D shape descriptor

Chapter 1 :

General Introduction

The picture is a clear message conveying civilization through the ages and generations, Since the Stone Age, Neanderthals were drawing and engraving on stone many coded messages that we understand today after the process of accurate interpretation and description of the content of those images.

Images is the best means of expression and communication despite the differences that exist between societies (Languages and concepts)

With the electronic development and computer science that we are witnessing today, we find it necessary to have a system of indexing and searching for images or for correct expression the need to have systems for making good use of images efficiently, easily and automatically.

This is the reason why the subject of image research has become an active topic in the world over the past few decades, which we consider to be a huge challenge in computer vision science.

A lot of effort has been put into researching image interpretation, but there is still no universally accepted approach to map low-level feature into high-level image semantic interpretation [1].

After that topic has made tremendous progress over the past years. The researchers focused on discovering and appreciating the concepts and objects in the images with high accuracy, as one of the proposed solutions to provide sufficient information to describe and interpret the semantic content of images and videos, however, the existence of differences in viewpoints form great difficulties that impede reaching the required level in the semantic interpretation of pictures.

This idea has received with great success in many applications, including face recognition, self-driving, and analysis of human behavior ... etc.

These applications rely on various techniques to obtain a suitable interpretation of what is inside the images for example, self-leadership depends primarily on the detection of objects.

With crimes on the rise all around the world, we thought video surveillance is becoming more important day by day. Due to the lack of human resources to monitor this increasing number of cameras manually, new computer vision algorithms to perform lower and higher-level tasks are being developed. In this thesis, we will discuss these algorithms and try to train them to discover humans and identify if humans are exposed to danger, and need help.

1. Definition of the interpretation and semantic indexing of an image content:

1.1. Image Interpretation :

The interpretation in this field is the process of examining multimedia documents like an image (such as aerial photo, medical image or digital image), a video for sensing image and identifying the features in that image. This method can be identify a wide variety of features, such as type and condition of vegetation on the riverside and human features …etc.

Image interpretation is based on seven characteristics that are inherent in imagery (also called image attributes) that we use to derive information about objects in an image .which are size, shape, tone/color, texture, shadow, association, and pattern.

The interpretation of data is challenging both in terms of automatically perceiving the images, extracting, and understanding differences between screened compounds and in terms of visualizing the results.

1.2. Semantic indexing:

Semantic indexing is a mathematical way of determining the relationship between various terms, topics and concepts in content. [5]
The indexing of images is based on the visual content of the image such as color, texture, shape, etc.

While these method are robust and effective they are still bottlenecked by semantic gap. That is, there is a significant gap between the high-level concepts (which human perceives) and the low-level features (which are used in describing images).

Semantic indexing of multimedia documents is generally carried out by detecting visual concepts via machine learning approaches supervised.[3]

2. Difficult and challenge of semantic indexing :

The existence of differences in viewpoints form great difficulties that impede reaching the required level in the semantic interpretation of pictures, these problems called "semantic gap", it is defined as follows.

2.1.The semantic and sensory gaps:

The semantic gap has typically been defined as the difference between the user's understanding of objects in an image and the computer's interpretation of those objects [2],
In addition to this, each user will interpret images differently and use different terms to label the objects within them or, more precisely the semantic gap characterizes the difference between two descriptions of an object by different linguistic representations [6].

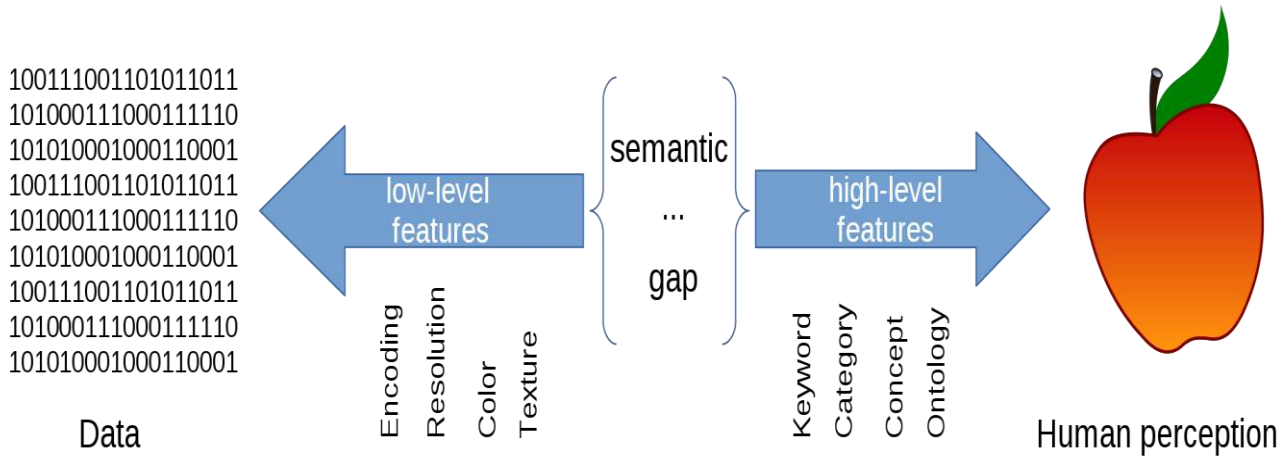


Figure 1: Semantic gap in the context of image analysis

There is another type of gap called "**sensory gap**", which is defined as the gap between the real world 3D and its representation in a 2D image. [3] On the other hand the **sensory gap** has been defined as the gap between a real life scene, and the information of this scene captured by sensors [2].

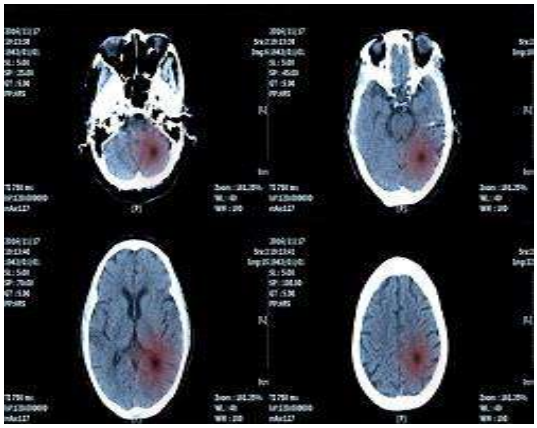
3. Application and Importance of Interpretation and semantic indexing of an image content :

The importance of interpretation and semantic indexing of image content is limited to the importance of accurate computer vision, which is considered a multidisciplinary scientific field Computer vision tasks include methods of how computers gain the ability to understand and extraction high-level data from digital images or videos. The interpretation and semantic indexing of image content is currently used in many fields like fashion ecommerce, and the beauty industry. The most prominent applications that use this feature are found in medicine, Military, industry, and Smart agriculture...Etc.

3.1. Industry field: In industry the interpretation and semantic indexing of image content is used for the purpose of supporting the manufacturing process, example (figure 2 (C)) quality control where details or final products are being automatically inspected in order to find defects.

3.2. Military field: In military, the obvious examples (figure 2 (B)) are detection of enemy soldiers or vehicles and missile guidance. More advanced systems for missile guidance send the missile to an area rather than a specific target, and target selection is made when the missile reaches the area based on locally acquired image data. [4]

3.3. Medicine field: In medicine an example (figure 2 (A)) of this is detection of tumors , arteriosclerosis or other malign changes; measurements of organ dimensions, blood flow.



(A) detect brain tumor

(B) advanced systems of military

(C) Machine vision quality control

Figure 2.1: Applications using interpretation and semantic indexing of image content

3.4. **Security:** for now the researchers are able to use object detection to identify anomalies in a scene such as bombs or explosives by making use of a quad copter.[7]

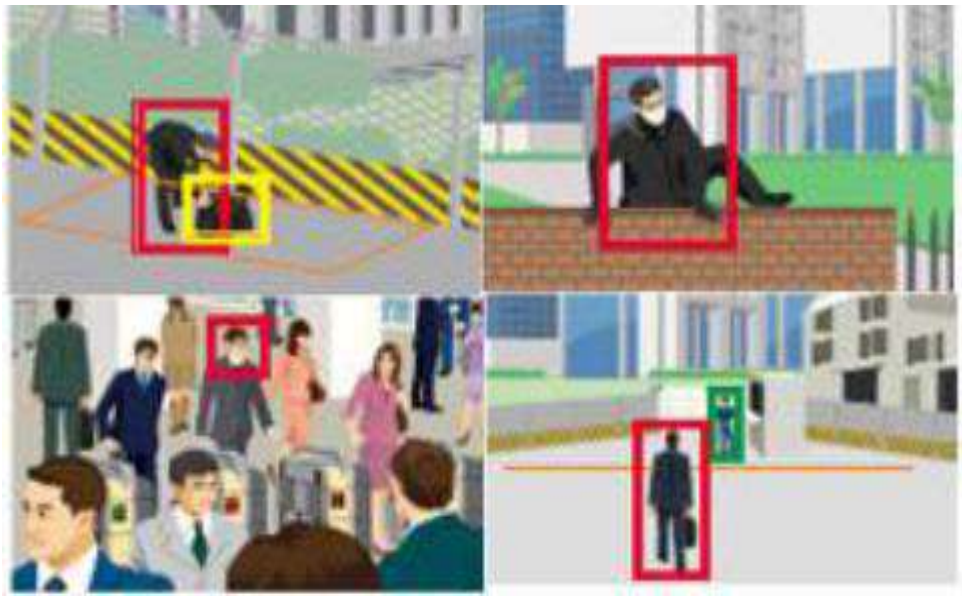


Figure 2. 2 : An example of security detection.

Other applications:

Figure 3.1: Customer has her face scanned by a face recognition system supported by face recognition technology of Alipay, the online payment service.

Figure 3.2: Smart farming, Farmer using NIR images and drones application screen used to check health alert cards for vegeta disease.

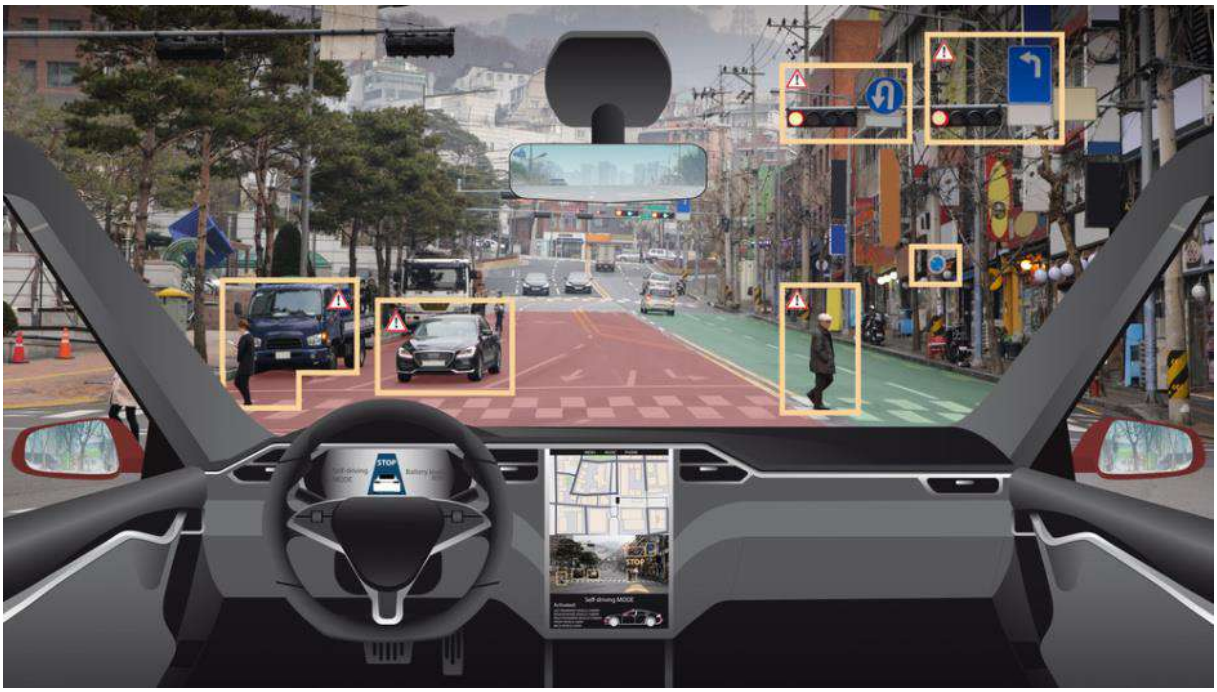


Figure 3.3: Self-driving car detect objects and interpret it

4. Our objective and motivation :

Reaching a high level of interpretation and semantic indexing of image content requires a lot of effort, and given its great importance in developing and facilitating human life today, this is what makes us contribute to this challenge.

We will try to detect the human in images or videos (i.e., series of images) that are plays an important role in various real life applications (e.g., visual surveillance and automated driver assistance).

Our objective in this thesis is train algorithms to recognize the human being at risk in order to develop surveillance cameras and Make it more accurate, and monitor human beings in order to preserve people's lives, combat crimes, and control civil laws to combat human violations against the environment for example, This is by using techniques and algorithms dedicated to dealing with images like CNN, YOLO, KNN.

5. Thesis organization :

Including this general introduction, the rest of this thesis is divided into three chapter are structured as follows:

- The second chapter we will take a look about the state of the art in domain of machine learning and will review the most important algorithms for classification and objects detection.
- In the third chapter, we will present the deep learning for human detection. We will start by introdus what is deep learning after that we will explaining methods in details how can detect a human, then we explain our methodology.
- The fourth chapter we will describe the experiment steps and evaluation, then the used metrics and the obtained results.
- General conclusion, which draws the conclusion of thesis, as it, illustrates the main outcome of it and what more can be achieved in the future.

Chapter 2:

State of the art

1. Introduction :

In order to learn how to extract information from images and process it automatically, to be used to solve many life problems, we need to clarify some of the concepts that are necessary about the main basics for semantic interpretation of image content in this chapter. In the first section, we will show the comparison and the relationship between image classification and object detection. Next, in the second section we will learn the process of extracting descriptors and features, in the third section we are going to describe the methods of machine learning and present the simple explanation of the different methods, Finally, the conclusion of this chapter.

2. Image classification vs. object detection in interpretation and semantic indexing of image content :

In general, image classification is only classify an image into a certain category , similar to object detection is identify the location of objects in an image, and e.g. count the number of instances of an object .

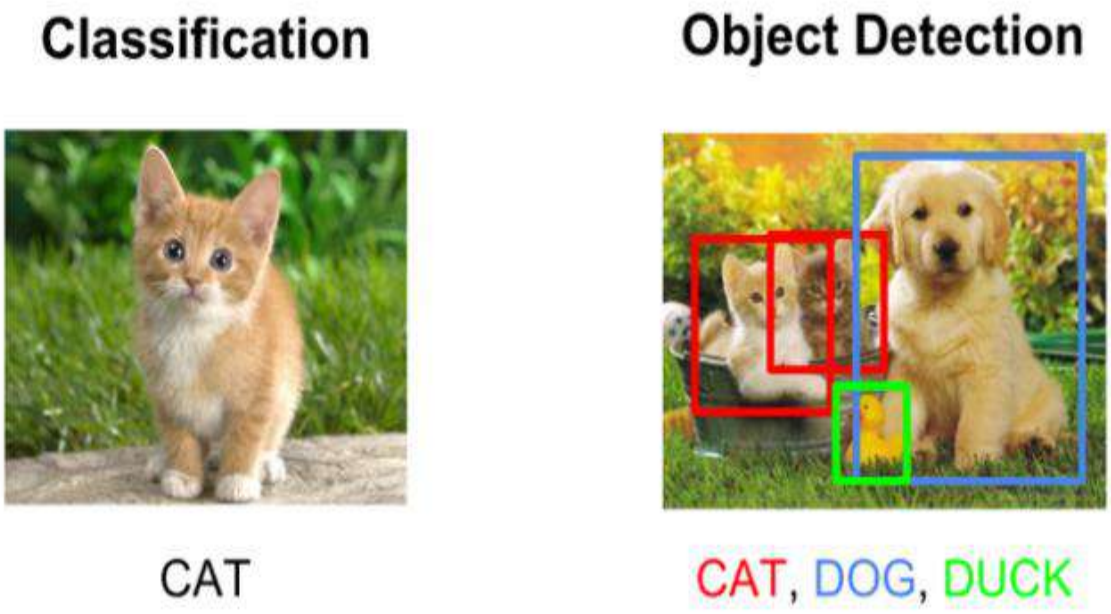


Figure 4: Image classification vs. object detection

2.1. Object detection :

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos [8] , Where it aims at locating and classifying existing objects in any given image, and labeling them with bounding boxes [9], It is widely used in computer vision tasks such as image annotation, activity recognition, face detection, face recognition, video object co-segmentation. It is also used in tracking objects, for example tracking a ball during a football match, tracking movement of a cricket bat, or tracking a person in a video.

2.2.Image classification:

Image classification is the primary domain, in which deep neural networks play the most important role of medical image analysis. The image classification accepts the given input images and produces output classification for identifying whether the disease is present or not. [10]

The object detection task is closely related to the image classification one, and gives information to interpret and analyze what is in the image content more than image classification.

3. Descriptors of an image :

The descriptors are the first step of computer vision, are generally used as input for algorithms. We can also categorize the descriptors according to the type of modality they represent: visual descriptors (they contain low-level descriptors which give a description about color, shape, regions, textures… etc.), audio descriptors, motion descriptors.
Those visual descriptors are presented by category in the Table 1 as follow.

Table 1: Visual descriptors

Visual descriptors	
Color	RGB DCD
Texture	TBD HTD
Shape	RSD 3-D SD
Feature (edge, corner ...)	HOG STIF

3.1.Feature descriptors:

It is a simplified representation of the image that contains only the most important information about the image. [5]

3.2.Feature extraction:

The basic principle of feature extraction is the detection of features from low to high levels. Low-level features include edges and colors. A high-level feature is an object, such as a cat, a tree... etc. [18]

The figure below explains this process with an example illustration.

(1) The first image had a lot of information, like the shape of the object, its color, the edges, background...Etc.

(2) The second had much less information (only the shape and the edges).

Figure 5: feature extraction

There are a number of feature descriptors, the most popular ones:

3.3.1. Histogram of oriented gradients(HOG):

HOG, or Histogram of Oriented Gradients, is a feature descriptor that is often used to extract features from image data. It is widely used in computer vision tasks for object detection. [11] The HOG descriptor focuses on the structure or the shape of an object.

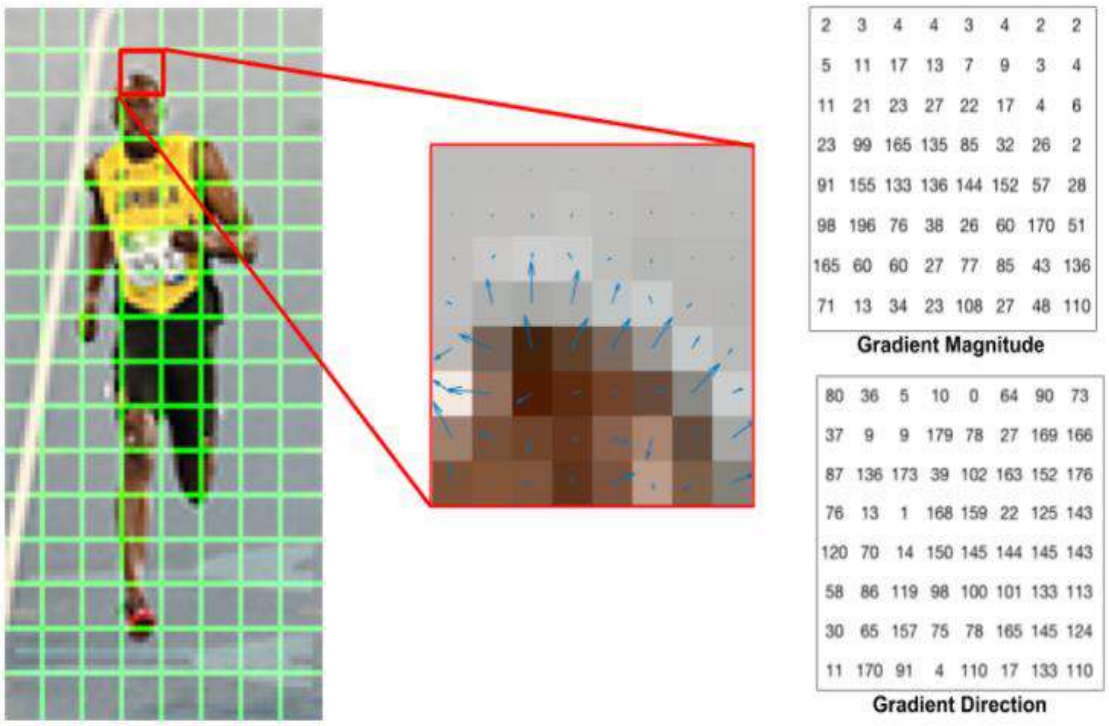


Figure 6: how to calculate the HOG feature vector.

Process of Calculating the Histogram of Oriented Gradients (HOG):

Step 1: resized the images to 64 x 128, this is to facilitate the process of dividing it into 8*8 and 16*16 patches to extract the features.

Step 2: Calculating Gradients (direction x and y)

Direction (G_x): subtract the value on the left from the pixel value on the right the selected pixel.
Direction (G_y): subtract the pixel value below from the pixel value above the selected pixel.
This process will give two new matrices one storing gradients in the x-direction and the other storing gradients in the y direction.

Step 3: Calculate the Magnitude and Orientation

Determine the magnitude and direction (orientation) for each pixel value:

Total Gradient Magnitude (the Pythagoras theorem) = $\sqrt{[(G_x)^2 + (G_y)^2]}$.

Orientation = Φ = atan (Gy / Gx)

Step 4: Calculate Histogram of Gradients in 8×8 cells (9×1)

Step 5: Normalize gradients in 16×16 cell (36×1)

V = [a1, a2, a3 ….a36]

We calculate the root of the sum of squares:

k = √ (a1)2+ (a2)2+ (a3)2+ …. (a36)2

And divide all the values in the vector V with this value k:

$$\text{Normalised Vector} = \left(\frac{a_1}{k}, \frac{a_2}{k}, \frac{a_3}{k}, \ldots \frac{a_{36}}{k}\right)$$

3.3.2. Scale Invariant Feature Transform (SIFT) :

The SIFT is a feature detection algorithm in computer vision to detect and describe local features in images. It was published by David Lowe in 1999 [12]. The SIFT descriptors are shown to be scale invariant and rotation and robust to noise and change in illumination. The works carried out within the framework of the indexing of multimedia documents showed that these descriptors give practically the best results, despite their large dimension.[3]

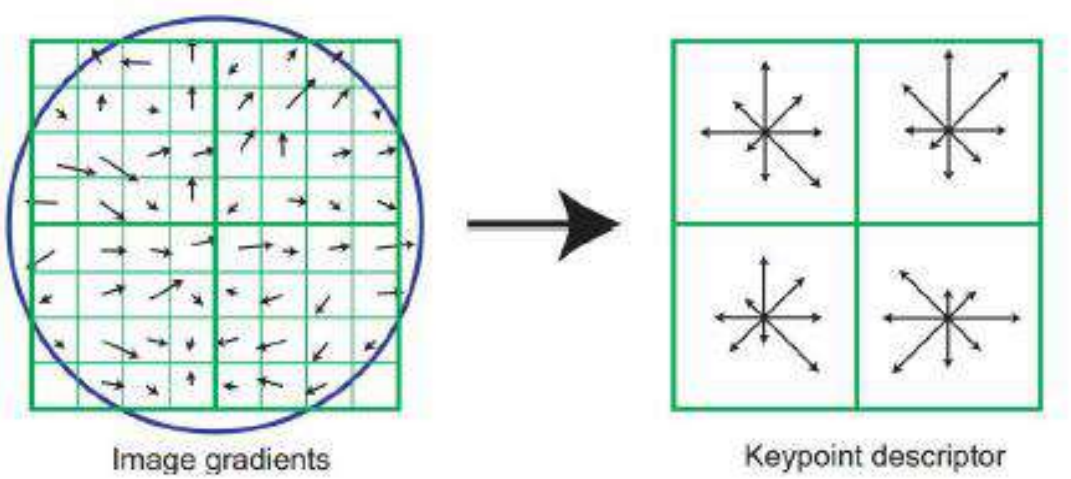


Figure 7: calculation principle of SIFT descriptors.

4. Machine learning :

Is a subset of artificial intelligence (AI) that uses mathematical and statistical approaches to give computers the ability to imitate human intelligence that is focused on creating systems that learn or improve performance based on the data they process. Machine Learning can be applied to different types of data, such as graphs, trees, multimedia documents, or more simply feature vectors, which can be continuous or discrete.

4.1. Types of machine learning :

Machine learning is divided in different ways depending on the learning mode they use , If the data is labeled it is a supervised learning , if the data is unlabeled we seek to define the underlying data structure (which can be probability density), then this is unsupervised learning. If the labels are discrete, or regression if they are continuous. If the model is learned incrementally based on a reward received by the program for each of the actions undertaken, we speak of reinforcement learning.

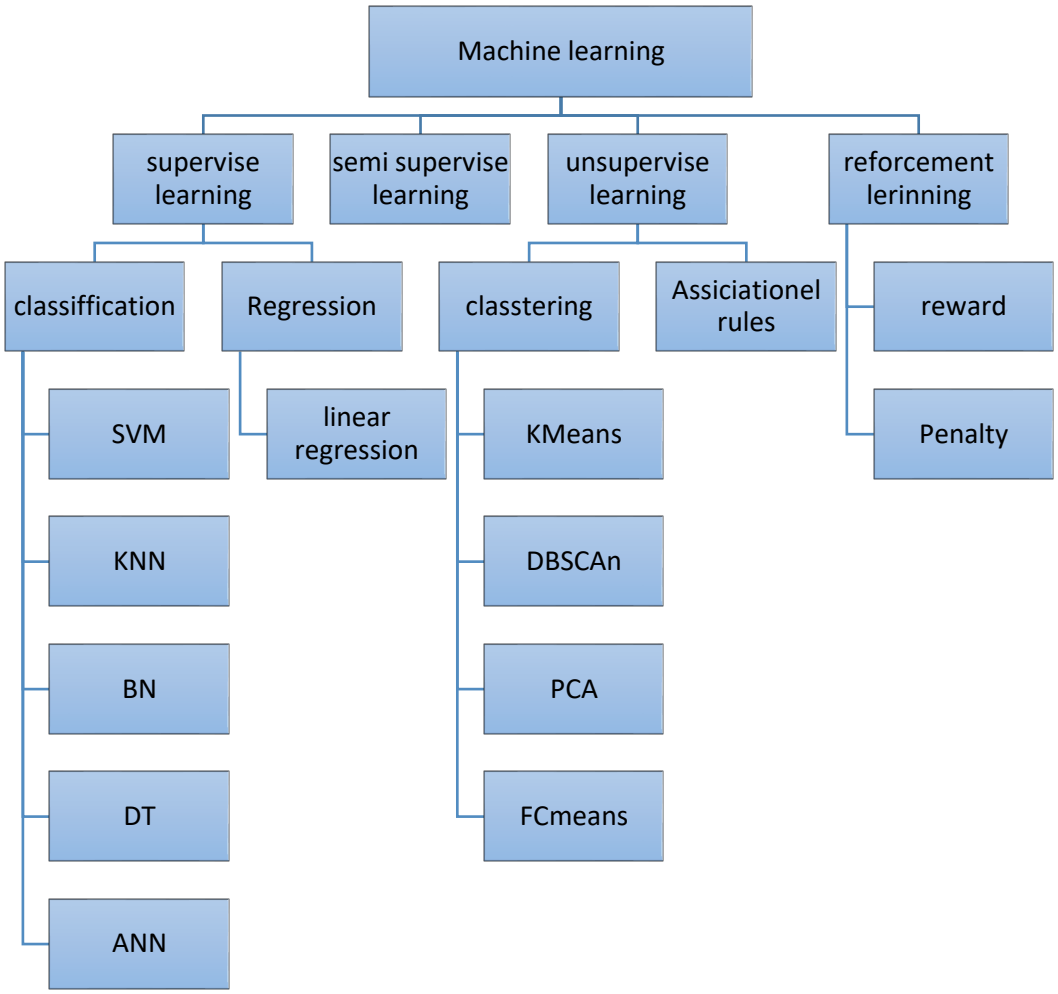


Figure 8: Types machine learning

4.1.1. Supervised Learning :

Supervised learning algorithms are the most commonly used. A supervised learning algorithm analyzes the training data and produces an inferred function, in this approach, the process are takes place in two phases. During the first phase (offline, so-called learning), this involves determining a model from the labeled data. The second phase (online, called test) consists in predicting the label of new data, knowing the previously learned model.

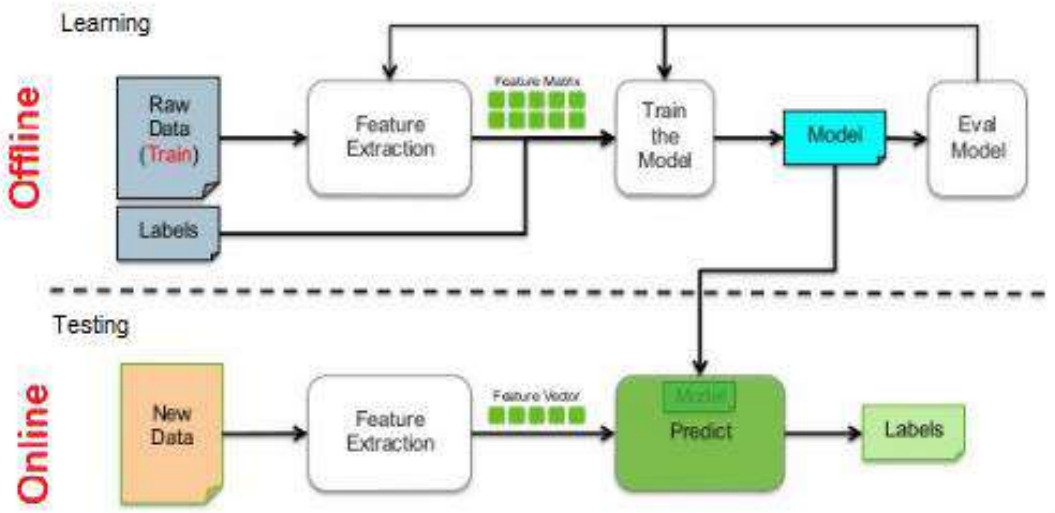


Figure 8.1: Supervised machine learning model

There are many supervised learning algorithms such as Support Vector Machines (SVM), k-nearest neighbors (KNN) and CNN.

4.1.1.1. k-nearest neighbors :

K-Nearest Neighbor (KNN) is one of the most popular machine learning algorithms can be used for both classification and regression predictive problems. Its operation can be compared to the following analogy "tell me who your neighbors are, I would tell you who you are ..."[14].
We can summarize the algorithm's steps as follows:

Step 1 − the first step of KNN, load the training as well as test data.

Step 2 − choose the value of K i.e. the nearest data points. K can be any integer.

Step 3 − For each point in the test data do the following:

Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
Step 4 – Sort the distance and determine the nearest neighbors based on k minimum distance.

Step 5 –The predicted class of the query-instance is affected based on the majority vote (The most frequent class) of the k closest neighbors selected in the previous step.

The figure above summarize all this steps of the K-Nearest Neighbor algorithm with an illustration example.

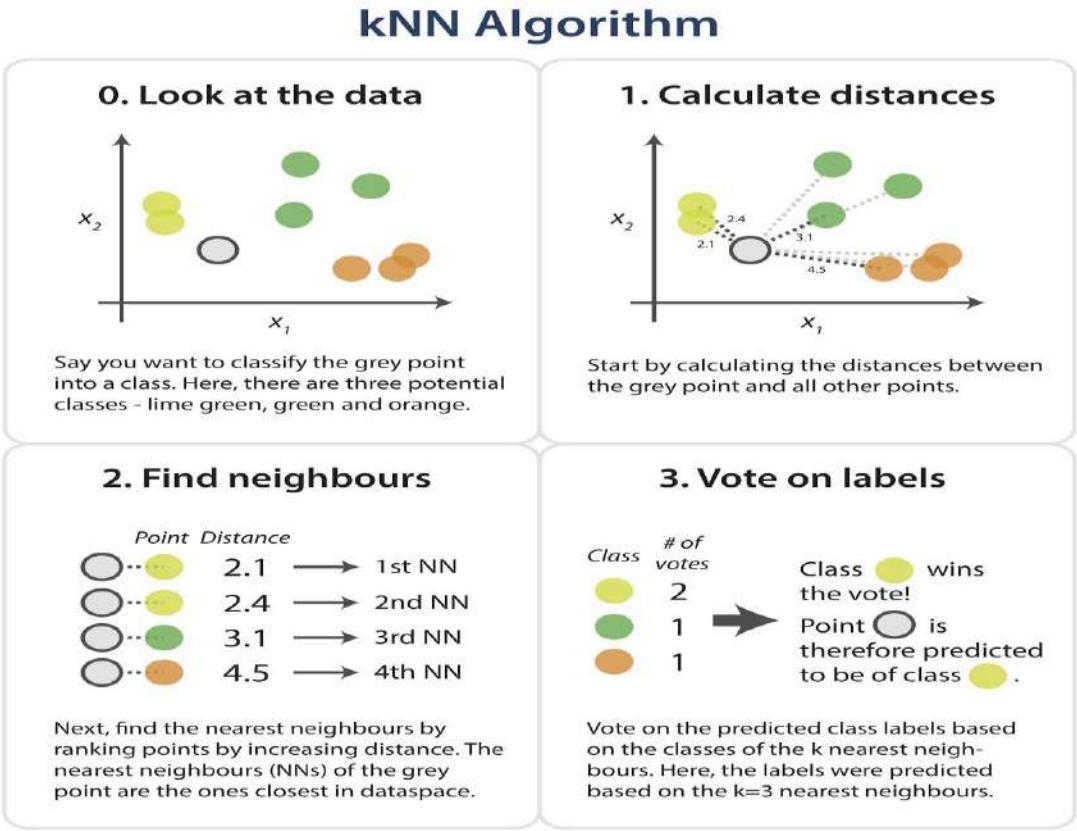


Figure 8.1.1: KNN algorithm steps

4.1.1.2. Support Vector Machine :

Support Vector Machine (SVM) is a supervised machine-learning algorithm , which can be used for both classification or regression. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. [13]
The principal of SVM algorithm is find the hyper-plane that differentiates the two classes.
There are two types of SVM can be linear or non-linear.

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, and then such data is termed as linearly separable data.

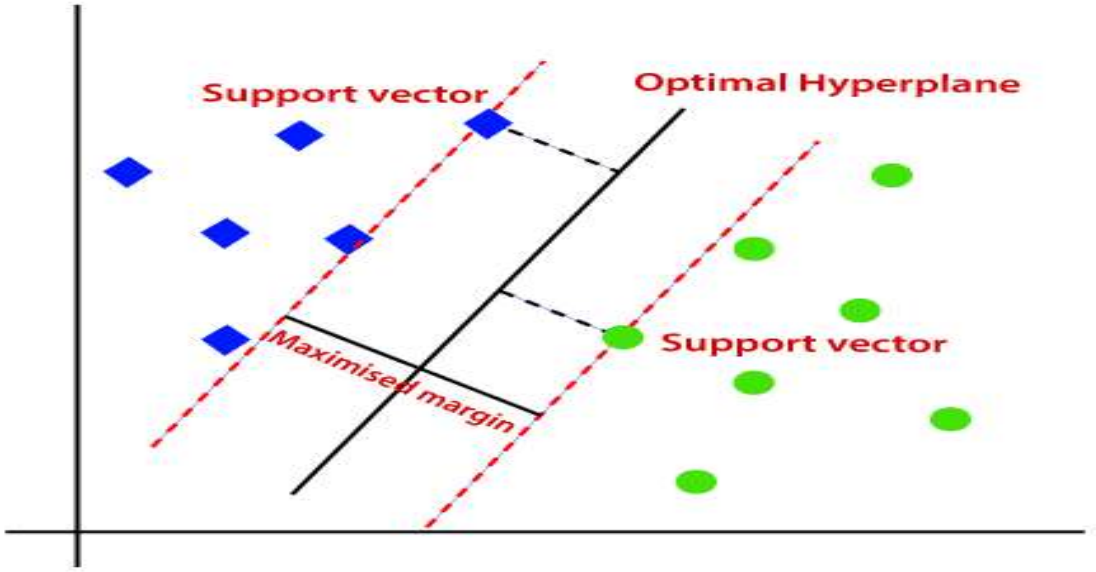


Figure 8.1.2: **Linear** SVM algorithm

Hyper-plane: is find the best line or decision boundary. The **hyper-plane** with maximum margin is called the **optimal hyper-plane**.
Support vectors: is the closest points of the lines from both the classes.
Margin: is the distance between the vectors and the hyper-plane and the goal of SVM is to maximize this margin.

- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, and then such data is termed as non-linear data. [15]

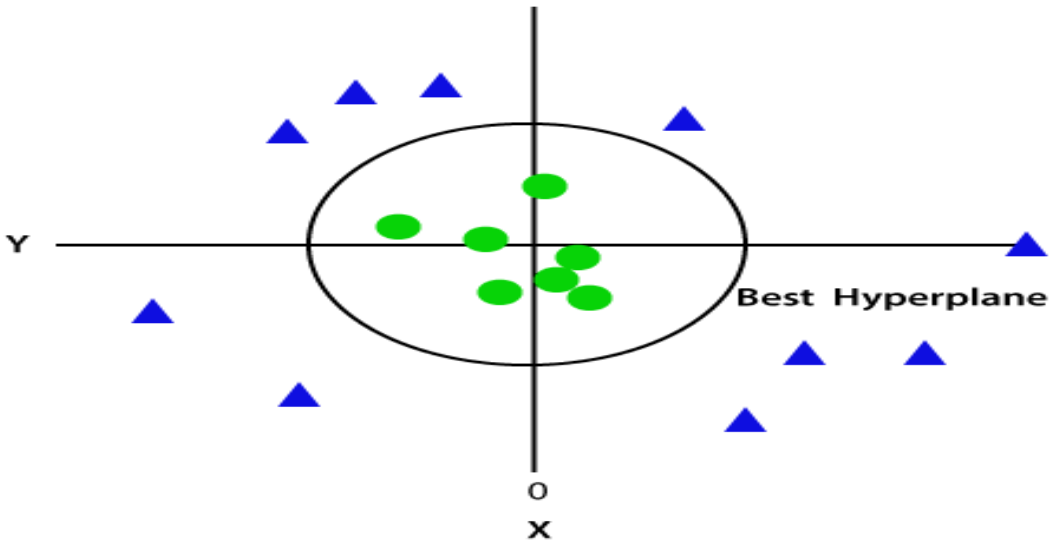


Figure 8.1.3: non-**Linear** SVM Algorithm

4.1.2. Unsupervised Learning (Clustering) :

Unsupervised learning involves training based on data without labels (the input data is unlabeled) or specific defined results, On the whole the goal is to *cluster* the data, to find reasonable groupings where the points in each group seem more similar to each other than to those in the other groups . Unlike classification, we are not aware of the types of clusters that will be formed at the end of the clustering algorithm. Several clustering algorithms were proposed like K-Means, FC-Means, and PCA.

The Following diagram depicts unsupervised learning model

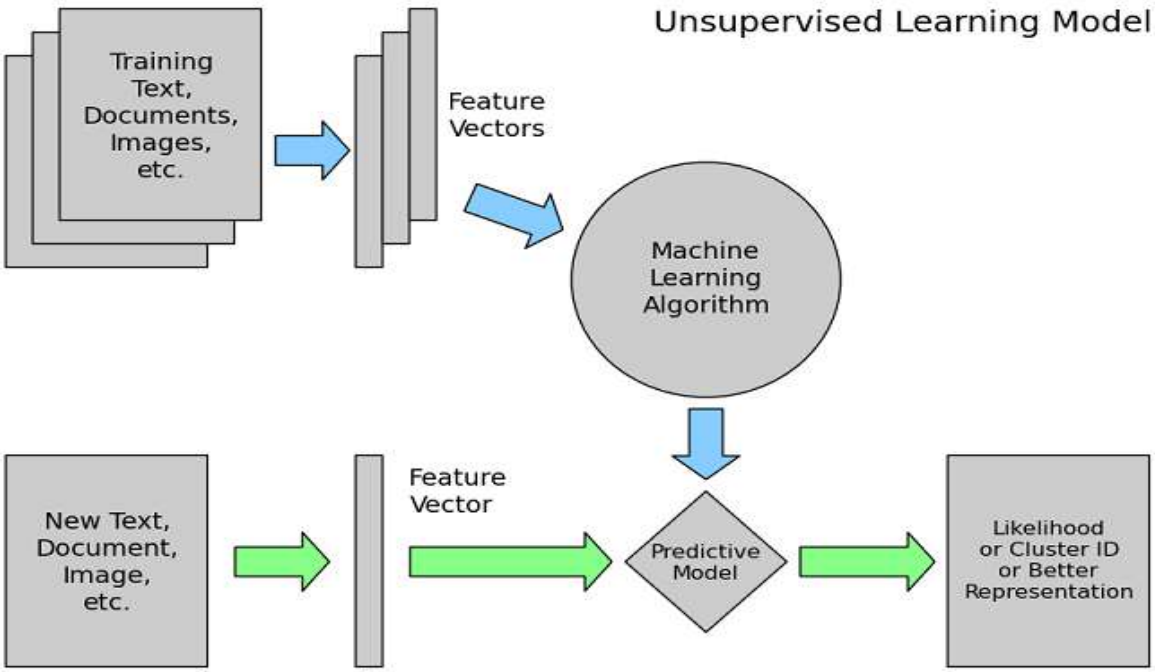


Figure 8.2: Unsupervised machine learning model

4.1.2.1. K-means clustering :

K-means clustering (Mac Queen, 1967) is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining them as follows:
1. Each instance di is assigned to its closest cluster center.
2. Each cluster center Cj is updated to be the mean of its constituent instances. [16]
If k is given, the K-means algorithm can be executed in the following steps:

- Partition of objects into k non-empty subsets
- Identifying the cluster centroids (mean point) of the current partition.
- Assigning each point to a specific cluster
- Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
- After re-allotting the points, find the centroid of the new cluster formed.[17]

The figure above summarize the steps of the k means algorithm with an illustration example.

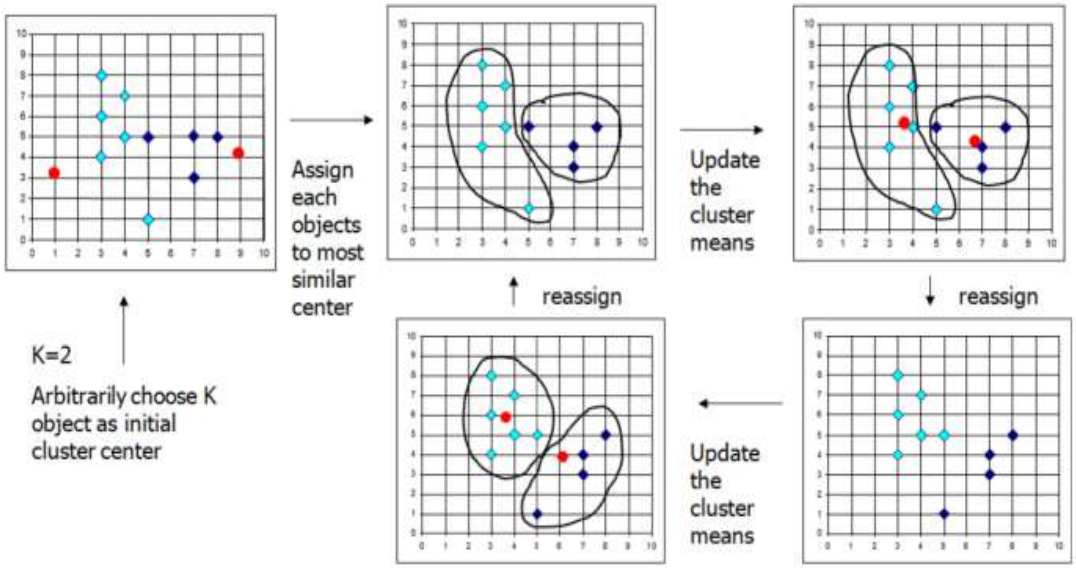


Figure 8.2.1: K-Means algorithm steps

5. Conclusion :

This chapter has been devoted to put the light into machine learning domain from different aspects (definition, types) and we have presented a global study on machine learning technique .This study includes both the feature extraction stage and classification phase are the two quite important steps in a machine learning.

In the next chapter, we will present deep learning for detecting a human and we will go in more details in very interested methods, which will be the base of our methodology.

Chapter 3:

Deep Learning

1. Introduction:

For domain of interpretation and semantic indexing of image content it exist many important applications which shows the important of this domain in real life like applications using techniques of objects detection .

In this chapter, we will focus especially on object detection techniques that are considered a still challenging task because, for the group of people, each individual has his unique appearance and, body shape.

First, we will explain the concept of deep learning for it is widely used in this field, after that we will present a set of algorithms for object detection, especially humans, and explain their basic steps. Then we move on to presenting a comparative study between the results of these techniques and their speed in detecting a human in images.

Finally, we touched on some deep learning classification techniques that we will rely on in our methodology

2. Deep Learning :

Deep learning is a type of artificial intelligence derived from machine learning (machine learning) where the machine is able to learn by itself . The concept of deep learning stems from research on artificial neural networks. Multiple-hidden layer perceptron is one of the structures of deep learning (Lecun, Bengio, & Hinton, 2015).

The main idea of depth learning is to incorporate low-level features into a more abstract level of advanced presentation attribute categories or features. The distributed characteristic representation of the data is found.

Deep learning is a kind of machine learning method based on data representation. Observations (such as images) can be expressed in a variety of ways. For example, a vector of the intensity value of each pixel, or more abstractly represented as a series of edges, areas of a particular shape, and so on (Ramachandran, Rajeev, Krishnan, & Subathra, 2015). [18]

2.1.How deep learning works :

Deep learning is based on a network of artificial neurons inspired by the human brain. This network is made up of tens or even hundreds of “layers” of neurons, each receiving and interpreting information from the previous layer. For example, the system will learn to recognize letters before tackling words in text, or determine if there is a face in a photo before finding out who it is.[19]

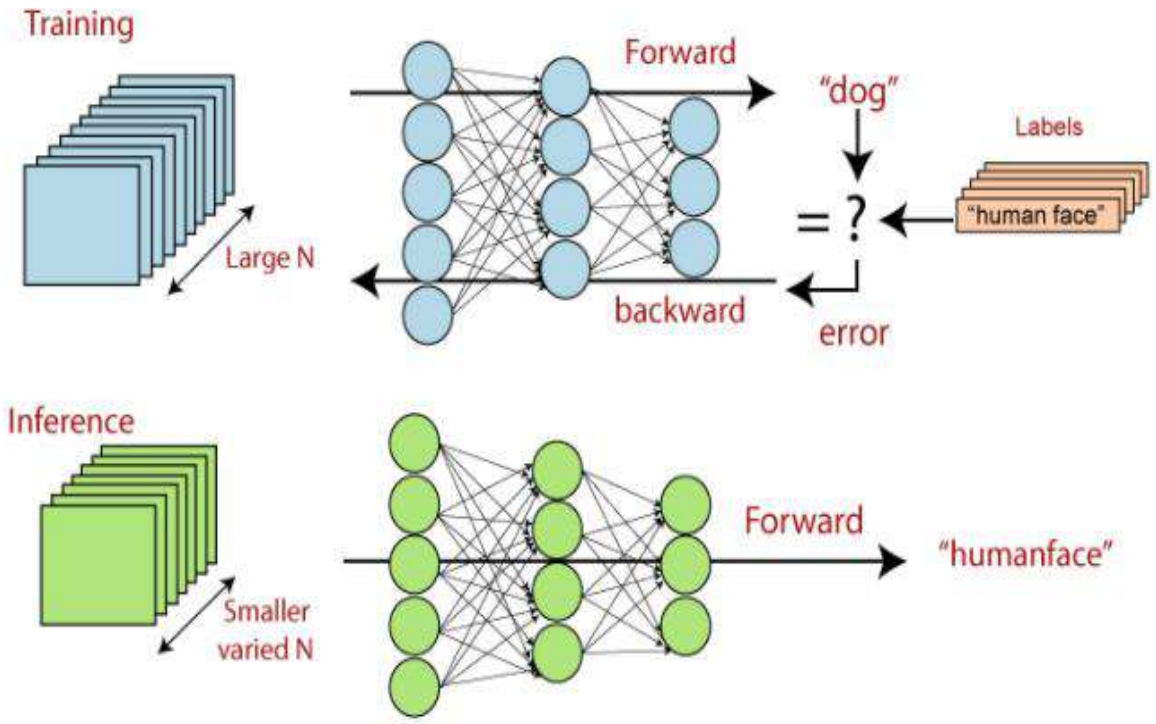


Figure 9: Diagram of deep learning.

In each step, the "wrong" answers are eliminated and sent back to the upstream levels to adjust the mathematical model. As it goes, the program reorganizes the information into more complex blocks. When this model is subsequently applied to other cases, he is normally able to recognize a cat without anyone ever telling him that he has never learned the concept of a cat. The baseline data is essential: the more the system accumulates different experiences, the better it will perform. [19]

2.2. Advantages of deep learning:

The deep learning does not require feature extraction manually, and it takes images directly as an input. It requires high-performance GPUs and lots of data for processing. The feature removal and classification are carried out by the deep learning algorithms this process is known as convolution neural network.

The performance of deep learning algorithms is improved when the amount of data increased. There are some advantages of deep learning which are given below:

1. The architecture of deep learning is flexible to be modified by new problems in the future.
2. The Robustness to natural variations in data is automatically learned.
3. Neural Network-based approach is applied in many different applications and data type.[20]

2.3.Disadvantages of deep learning:

There are several disadvantages of deep learning which are given below:

1. We require a very large amount of data in deep learning to perform better than other techniques.
2. There is no standard theory to guide you in selecting the right deep learning tools. This technology requires knowledge of topology, training methods, and other parameters. As a result, it is not simple to be implemented by less skilled people.
3. The deep learning is extremely expensive to train the complex data models.
4. It requires expensive GPUs and hundreds of machines, and this increases the cost of the user.[20]

3. Deep learning for object detection :

There are mainly two types of state-of-the-art object detectors. On one hand, we have two-stage detectors, such as Faster R-CNN (Region-based Convolutional Neural Networks) or Mask R-CNN, that (i) use a Region Proposal Network to generate regions of interests in the first stage and (ii) send the region proposals down the pipeline for object classification and bounding-box regression. Such models reach the highest accuracy rates, but are typically slower.

On the other hand, we have single-stage detectors, such as YOLO (You Only Look Once) and SSD (Singe Shot MultiBox Detector), that treat object detection as a simple regression problem by taking an input image and learning the class probabilities and bounding box coordinates. Such models reach lower accuracy rates, but are much faster than two-stage object detectors. [21]

The Figure (10.1) below is a popular example of illustrating how an object detection algorithm works. Each object in the image, from a person to a kite, have been located and identified with a certain level of precision.

The Figure (10.2) is an example of detecting person, each person in the image have been located and identified with a certain level of precision.

Figure 10.1: examples an Object Detection Task (using Deep Learning)

Figure 10.2: examples an person detection (using Yolo)

There are a lot of methods used in human detection domain the following Table 2 presented those methods category.

Table 2: categories of object detectors models

One-stage detectors	Two-stage detectors
YOLO SSD	R-CNN Fast R-CNN Faster R-CNN Mask R-CNN

3.1. Region-based Convolutional Network (R-CNN) :

To extract high-level features it is important to improve the quality of candidate bounding boxes and to get a deep structure. To solve these problems, R-CNN [23] was proposed by Ross Girshick in 2014 as an alternative exhaustive to capture object location in images. It starts to initializes small regions in the input image then merges them according to variety of color and similarity metrics using a hierarchical grouping as in Figure 9.

Thus, the final group is a box containing the whole image. The output is a few numbers of regions which could contain an object. [22]

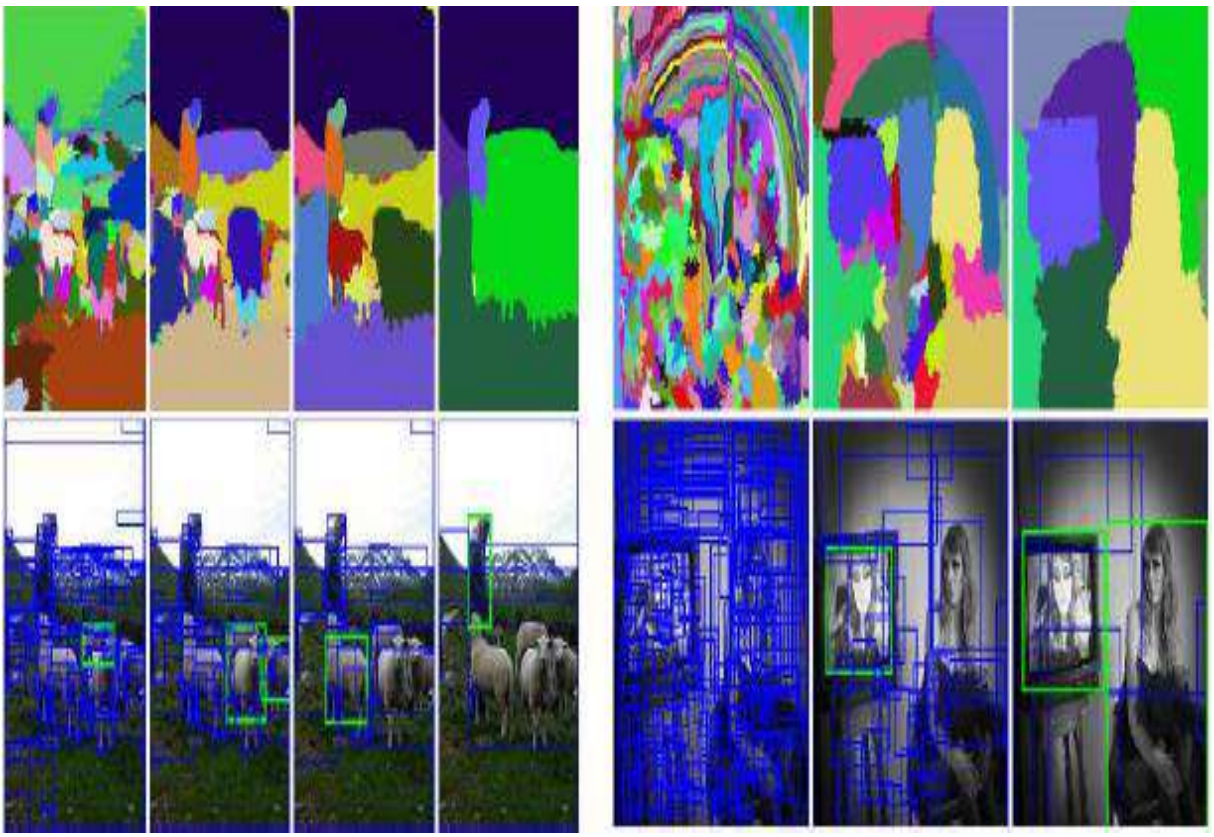


Figure 11: Two examples showing the result of the algorithm top: visualization of the segmentation results, down: visualization of the region proposals [24]

R-CNN, or Region-based Convolutional Neural Network, consisted of 3 simple steps:

1. Scan the input image for possible objects using an algorithm called Selective Search, generating ~2000 region proposals
2. Run a convolutional neural net (CNN) on top of each of these region proposals
3. Take the output of each CNN and feed it into a) an SVM to classify the region and b) a linear regressor to tighten the bounding box of the object, if such an object exists.

The entire process of object detection using RCNN has three models:

- CNN for feature extraction
- Linear SVM classifier for identifying objects
- Regression model for tightening the bounding boxes.

These 3 steps are illustrated in the image below:

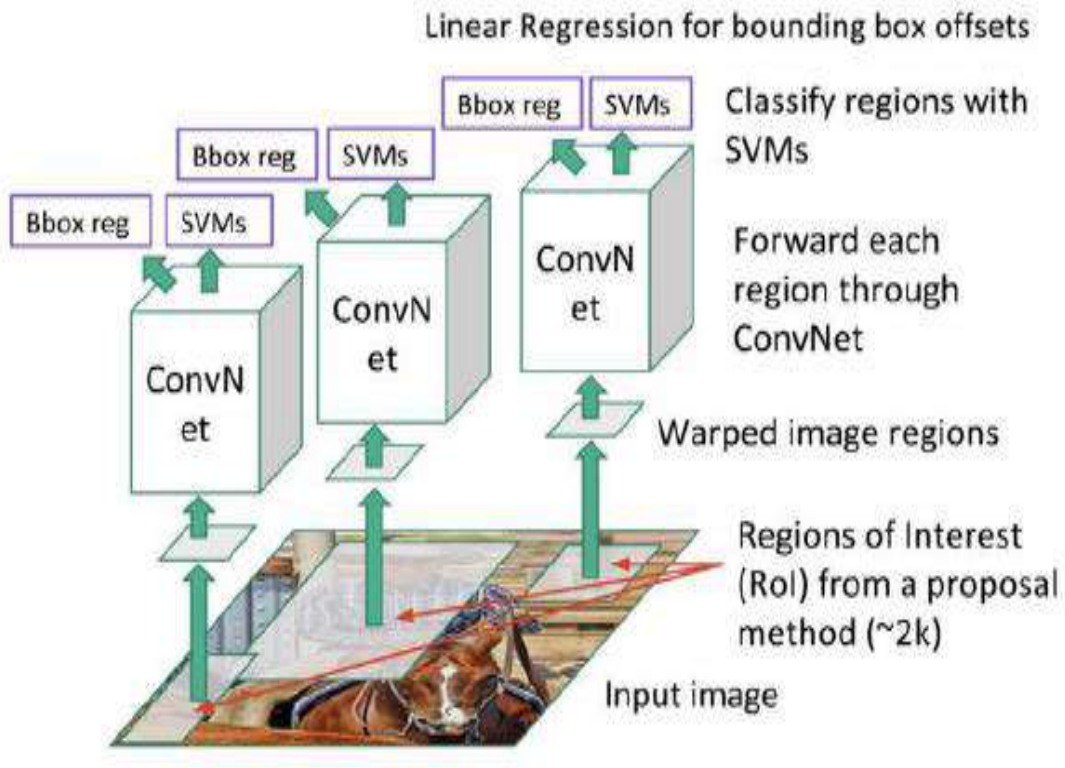


Figure 12: Region-based Convolution Network (R-CNN). Each region proposal feeds a CNN to extract a features vector, possible objects are detected using multiple SVM classifiers and a linear regression modifies the coordinates of the bounding box. [26]

In other words, we first propose regions, then extract features, and then classify those regions based on their features. In essence, we have turned object detection into an image classification problem. R-CNN was very intuitive, but very slow. [26]

3.2.Fast Region-based Convolutional Network (Fast R-CNN):

The goal of this method is to reduce the time, which related to the high number of models necessary to analyze all region proposals [27]. In Fast RCNN, we feed the input image to the CNN, which in turn generates the convolutional feature maps as in **Figure 13**. Using these maps, the regions of proposals are extracted. We then use a RoI pooling layer to reshape all the proposed regions into a fixed size, so that it can be fed into a fully connected network. [28] Fast RCNN resolves two major issues of RCNN, i.e., passing one instead of 2,000 regions per image to the ConvNet, and using one instead of three different models for extracting features, classification and generating bounding boxes. [28]

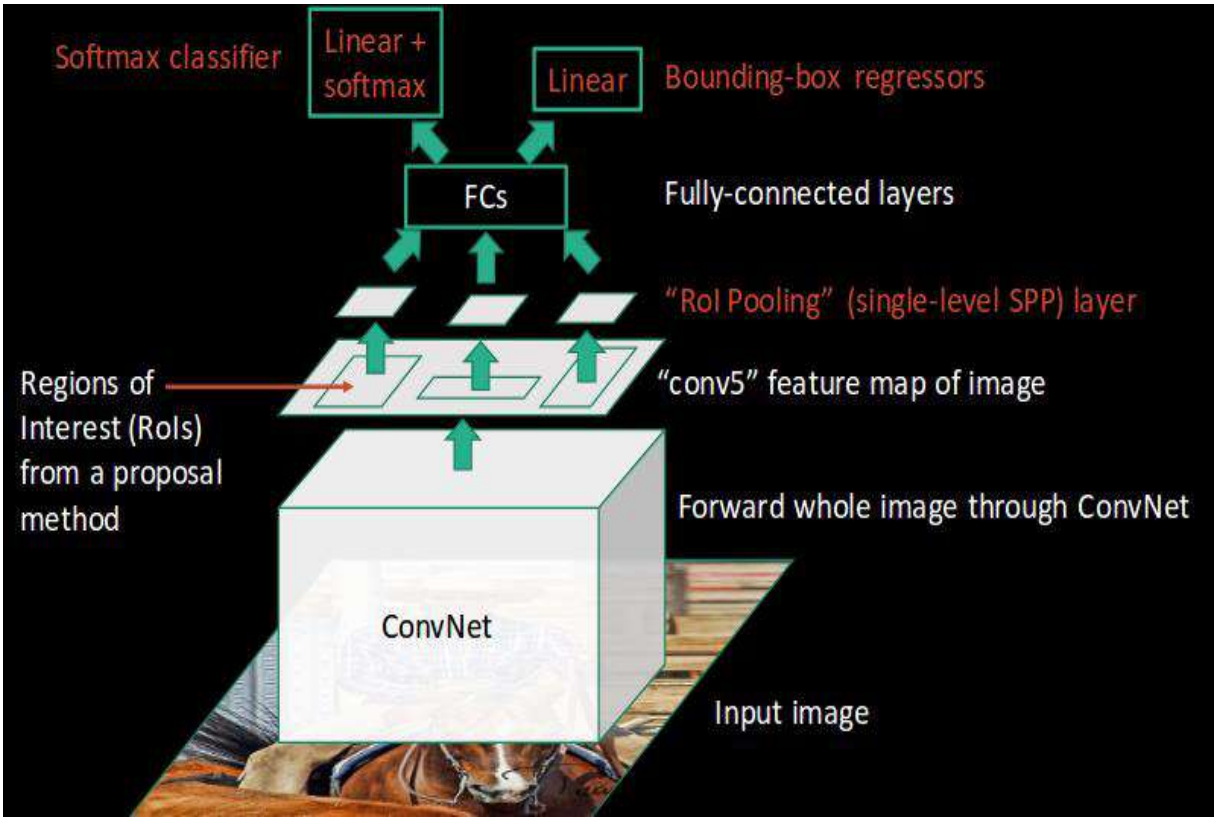


Figure 13: Architecture of Fast RCNN

Let's break this down into steps to simplify the concept:

- Passe the image input to a ConvNet, which in turns generates the Regions of Interest.
- A RoI pooling layer is applied on all of these regions to reshape them as per the input of the ConvNet. Then, each region is passed on to a fully connected network.
- A softmax layer is used on top of the fully connected network to output classes. Along with the softmax layer, a linear regression layer is also used parallely to output bounding box coordinates for predicted classes.

3.3.Faster Region-based Convolutional Network (Faster R-CNN):

This method is a combination between the Region Proposal Network (RPN) [29] and the Fast R-CNN model. [29] The entire image will be as input to the CNN model, after that the CNN model produces the feature maps. All the feature maps will slides by a window of size 3x3 and outputs a features vector linked to two fully connected layers, one for box-classification and the other one for box-regression. Multiple region proposals are predicted by the fully connected layers. A maximum of *k* regions is fixed thus the output of the box-regression layer has a size of 4*k* (coordinates of the boxes, their height and width) and the output of the box-classification layer a size of 2*k* (scores to detect an object or not in the box). The *k* region proposals detected by the sliding window are called anchors. [30]

The below steps are typically followed in a Faster RCNN approach [28]:

- Take an image as input and pass it to the ConvNet, which returns the feature map for that image.
- Region proposal network is applied on these feature maps. This returns the object proposals along with their objectness score.
- A RoI pooling layer is applied on these proposals to bring down all the proposals to the same size.
- Finally, the proposals are passed to a fully connected layer which has a softmax layer and a linear regression layer at its top, to classify and output the bounding boxes for objects.

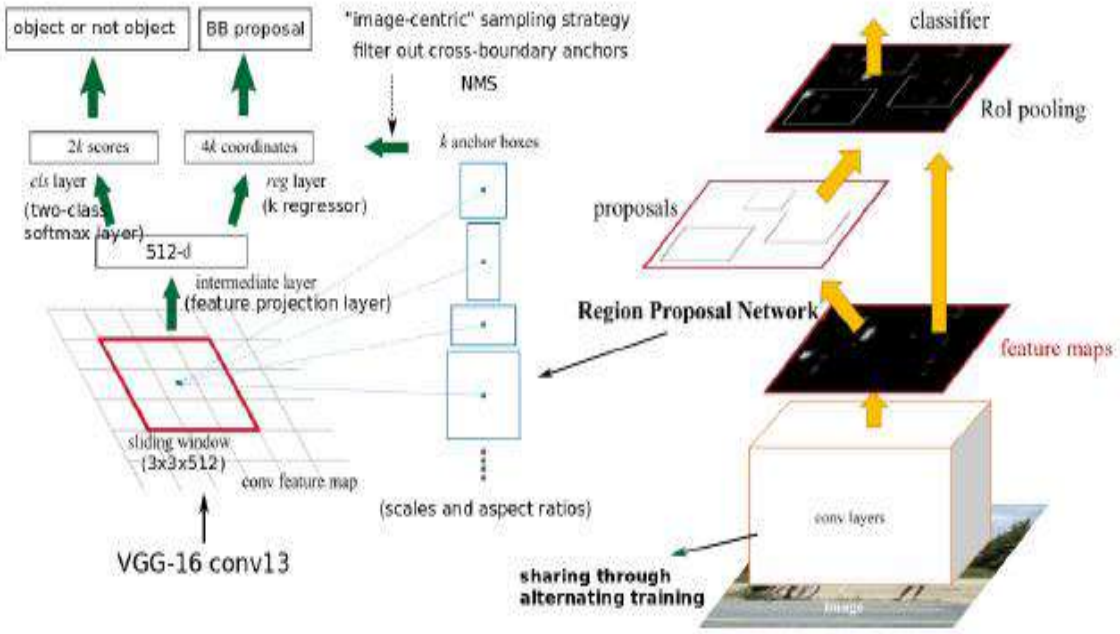


Figure 14: Architecture of Faster RCNN

The below table 3 is a comparative summary of some the family algorithms of RCNN [28]

Table 3: comparative summary of some the family algorithms of RCNN

Algorithm	Features	Prediction time / image
RCNN	Uses selective search to generate regions. Extracts around 2000 regions from each image.	40-50 seconds
Fast RCNN	Each image is passed only once to the CNN and feature maps are extracted. Selective search is used on these maps to generate predictions. Combines all the three models used in RCNN together.	2 seconds
Faster RCNN	Replaces the selective search method with region proposal network which made the algorithm much faster.	0.2 seconds

3.4. You Only Look Once (YOLO) :

This method is basically getting the image and split it into an *SxS* grid each one contain *m* bounding boxes. For this last the network outputs a class probability and offset values.

The bounding boxes having the class probability above a threshold value is selected and used to locate the object within the image. YOLO is classified as one of the faster (45 frames per second) object detection algorithms. The limitation of this method is that it finds difficulties with small objects in the image. [31]

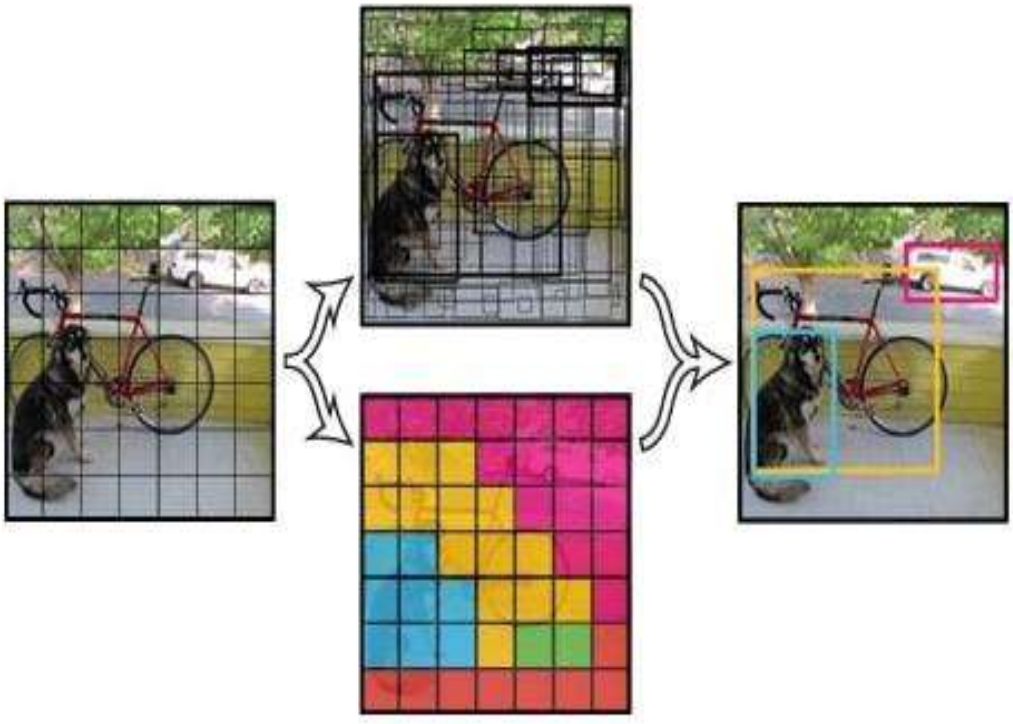


Figure 15: example objects detection using Yolo

Below are the exact dimensions and steps that the YOLO algorithm follows:

- Takes an input image of shape (608, 608, 3)
- Passes this image to a convolutional neural network (CNN), which returns a (19, 19, 5, 85) dimensional output
- The last two dimensions of the above output are flattened to get an output volume of (19, 19, 425):
    - Here, each cell of a 19 X 19 grid returns 425 numbers
    - 425 = 5 * 85, where 5 is the number of anchor boxes per grid
    - 85 = 5 + 80, where 5 is (pc, bx, by, bh, bw) and 80 is the number of classes we want to detect
- Finally, we do the IoU and Non-Max Suppression to avoid selecting overlapping boxes

3.5. The Single Shot Detector (SSD):

The model takes an image as input which passes through multiple convolutional layers with different sizes of filter (10x10, 5x5 and 3x3). To predict the bounding boxes Feature maps from convolutional layers at different position of the network are used. They are processed by specific convolutional layers with 3x3 filters called extra feature layers to produce a set of bounding boxes similar to the anchor boxes of the Fast R-CNN. [30]

In the figure below, the first few layers (white boxes) are the backbone, the last few layers (blue boxes) represent the SSD head.

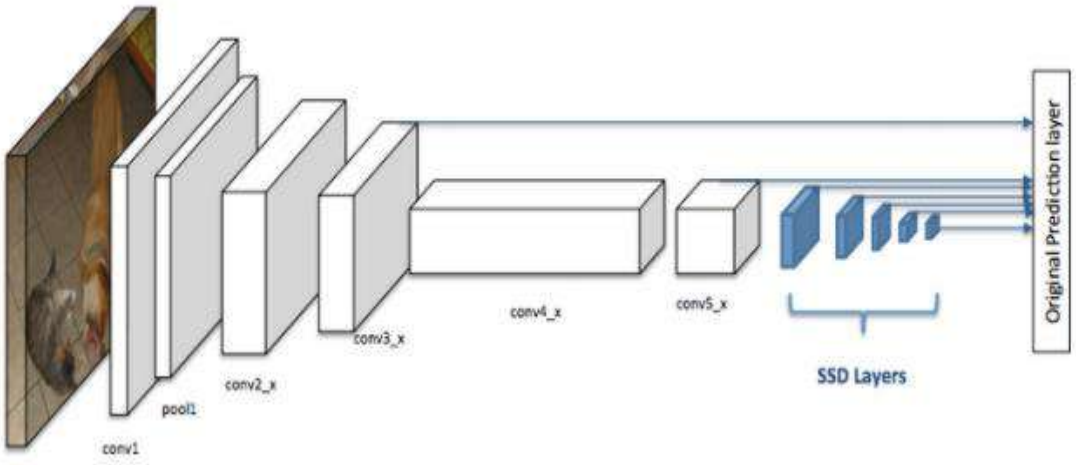


Figure 16: Architecture of a convolutional neural network with a SSD detector [29]

3.6.Evaluation :

The comparison will be between various object detection methods on two datasets, which are PASCAL VOC 2007, PASCAL VOC 2012.

The comparison will be between various object detection methods on two datasets, which are PASCAL VOC 2007, PASCAL VOC 2012.

1. **PASCAL VOC 2007/2012:** PASCAL VOC (Visual Object Classes) 2007[38] and 2012[39] datasets consist of 20 categories. The evaluation terms are Average Precision (AP) in each single category and mean Average Precision (mAP) across all the 20 categories.
2. **Microsoft COCO:** Microsoft COCO (Common objects in Context) [40]. Is composed of 300,000 fully segmented images, in which each image has an average of seven object instances from a total of 80 categories. As there are a lot of less iconic objects with a broad range of scales and a stricter requirement on object localization, this dataset is more challenging than PASCAL 2012. Object detection performance is evaluated by AP computed under different degrees of IOUs and on different object sizes.

Table 4: Real-Time Systems on PASCAL VOC 2007. Comparing the performance and speed of fast detectors [33]

Real-Time Detectors	Train	mAP	FPS
100Hz DPM [30]	2007	16.0	100
30Hz DPM [30]	2007	26.1	30
Fast YOLO	2007+2012	52.7	**155**
YOLO	2007+2012	**63.4**	45
Less Than Real-Time			
Fastest DPM [37]	2007	30.4	15
R-CNN Minus R [20]	2007	53.5	6
Fast R-CNN [14]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[27]	2007+2012	73.2	7
Faster R-CNN ZF [27]	2007+2012	62.1	18
YOLO VGG-16	2007+2012	66.4	21

The table below provides a comparison between the results of previous techniques in detecting humans in images trained on a dataset.

Table 5: Comparative results on VOC 07/12 and Microsoft COCO test set (%). [32]

Methods	Trained on			Person
	Coco	Voc 2007	Voc 2012	
R-CNN			✓	57,8
Fast R-CNN		✓	✓	72,0
Faster R-CNN	✓	✓	✓	84,1
YOLO		✓	✓	63,5
SSD	✓	✓	✓	85,6

4. Deep learning for Image classification :

Classification is a systematic arrangement in groups and categories based on its features. Image classification came into existence for decreasing the gap between the computer vision and human vision by training the computer with the data. The image classification is achieved by differentiating the image into the prescribed category based on the content of the vision.[43]

4.1.VGG16 :

VGG 16 was proposed by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group Lab of Oxford University in 2014 in the paper "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION". [44] It was one of the famous model using for classification and detection.

The figure below illustrates the architecture of this model.

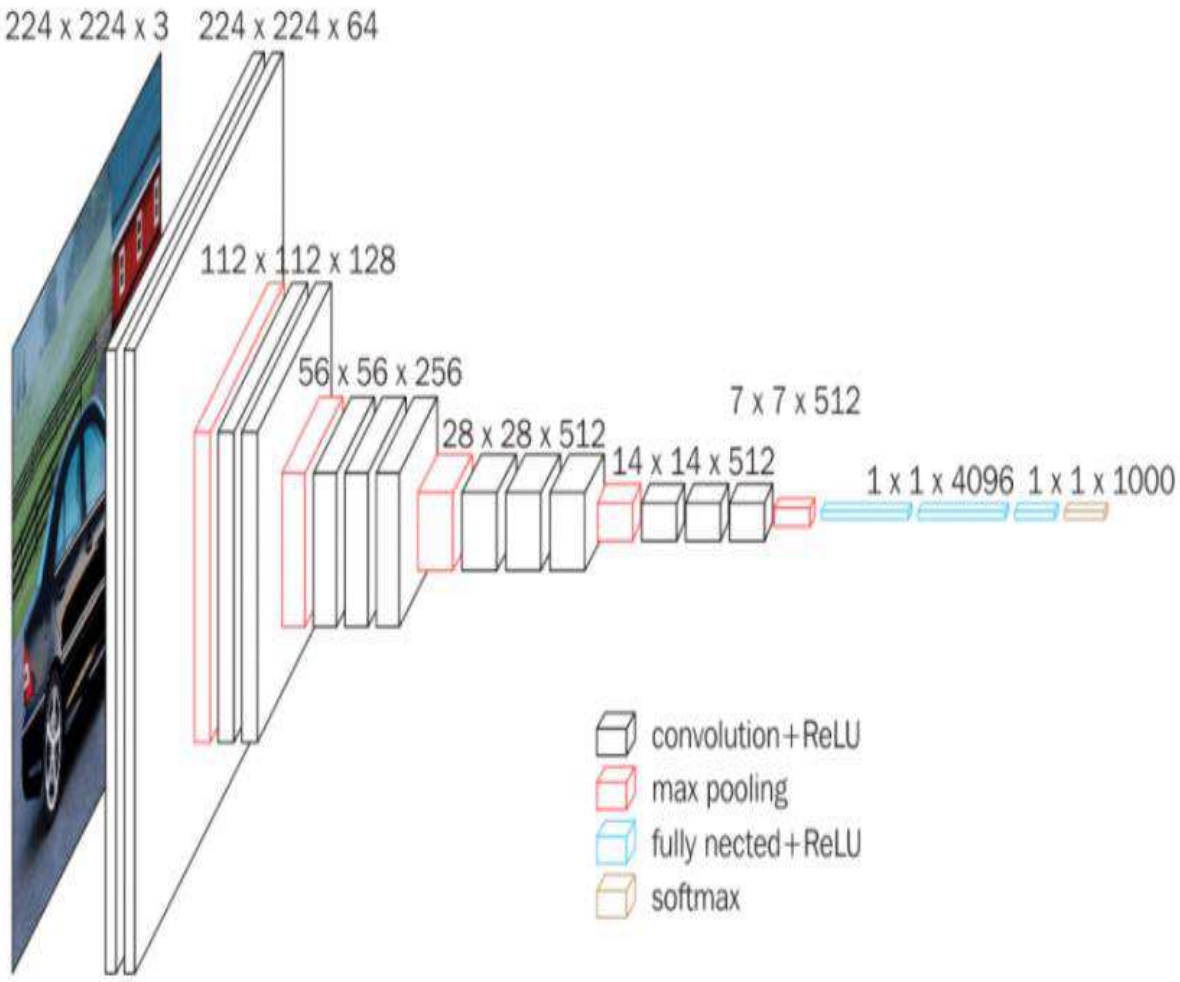


Figure 17: The architecture of VGG16.

In this figure:
The input to the network is image of dimensions *(224, 224, 3)*.
The first two layers have *64* channels of *3*3* filter size and same padding.
Then after a max pool layer of stride *(2, 2)*, two layers which have convolution layers of 256 filter size and filter size *(3, 3)*.
This followed by a max-pooling layer of stride *(2, 2)* which is same as previous layer.
Then there are *2* convolution layers of filter size *(3, 3)* and *256* filter.
After that there are *2* sets of *3* convolution layer and a max pool layer. Each have *512* filters of *(3, 3)* size with same padding.
This image is then passed to the stack of two convolution layers. In these convolution and max pooling layers, the filters we use is of the size *3*3* instead of *11*11* in Alex Net and *7*7* in ZF-Net.
In some of the layers, it also uses *1*1* pixel which is used to manipulate the number of input channels.
There is a padding of *1-pixel* (same padding) done after each convolution layer to prevent the spatial feature of the image. [44]

4.2. Inception V3 :

is a convolutional neural network architecture from the Inception family that makes several improvements including using Label Smoothing, Factorized 7 x 7 convolutions, and the use of an auxiliary classifer to propagate label information lower down the network (along with the use of batch normalization for layers in the side head).[45]

The figure below illustrates the architecture of this model.

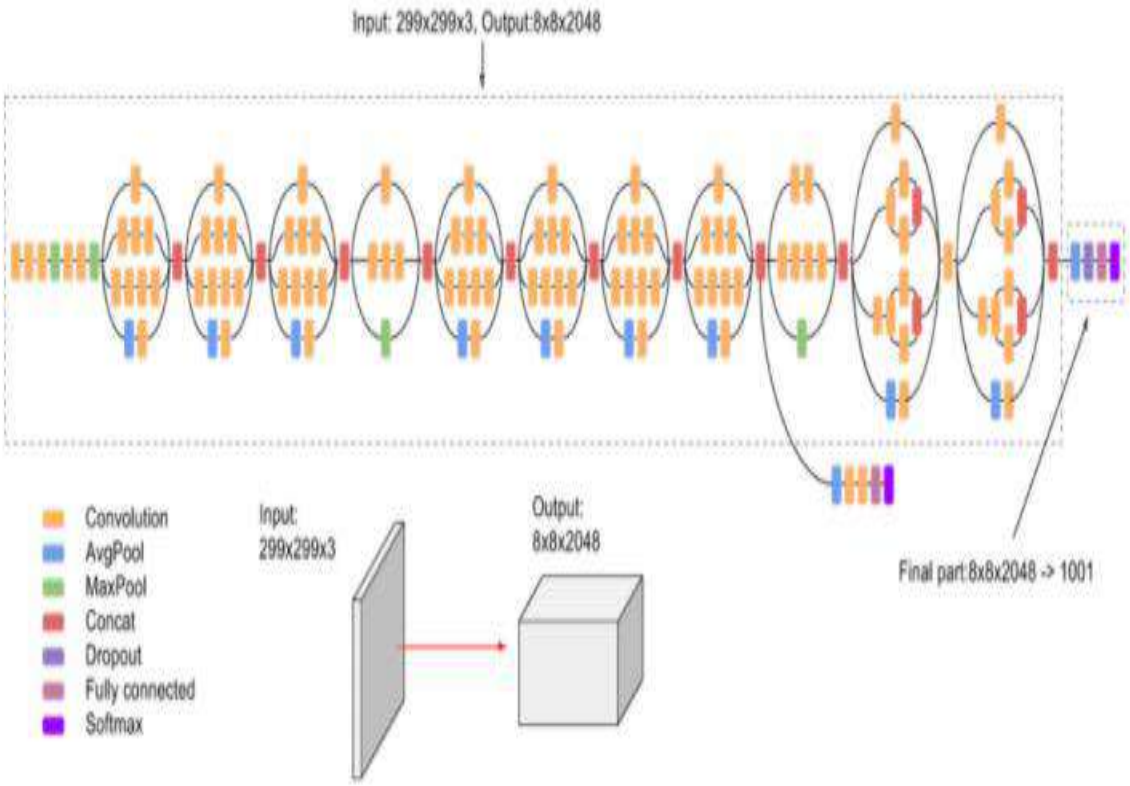


Figure 18: The architecture of Inceptionv3

6. Conclusion:

In this chapter we presented the main methods that are using for detect a human in image like YOLO and Fast RCNN. In addition to that, we explained the difference between these algorithms. We also presented some deep learning methods that are used in image classification like Vgg16. In the coming chapter we will show all the experiments that we did and all the used materials and the results that we get.

Chapter 4 :

Expriment results

1. Introduction :

In this chapter, we will present our methodology and how it works and we will present the experiment results and all the used materials, the dataset that we used the, the evaluation metrics and the results that we get and every small detail concern how we get the results.

2. Our methodology :

In this part we will explain our idea that we follow to get our result,

Our methodology is based on image classification and object detection together in the figure19 below.

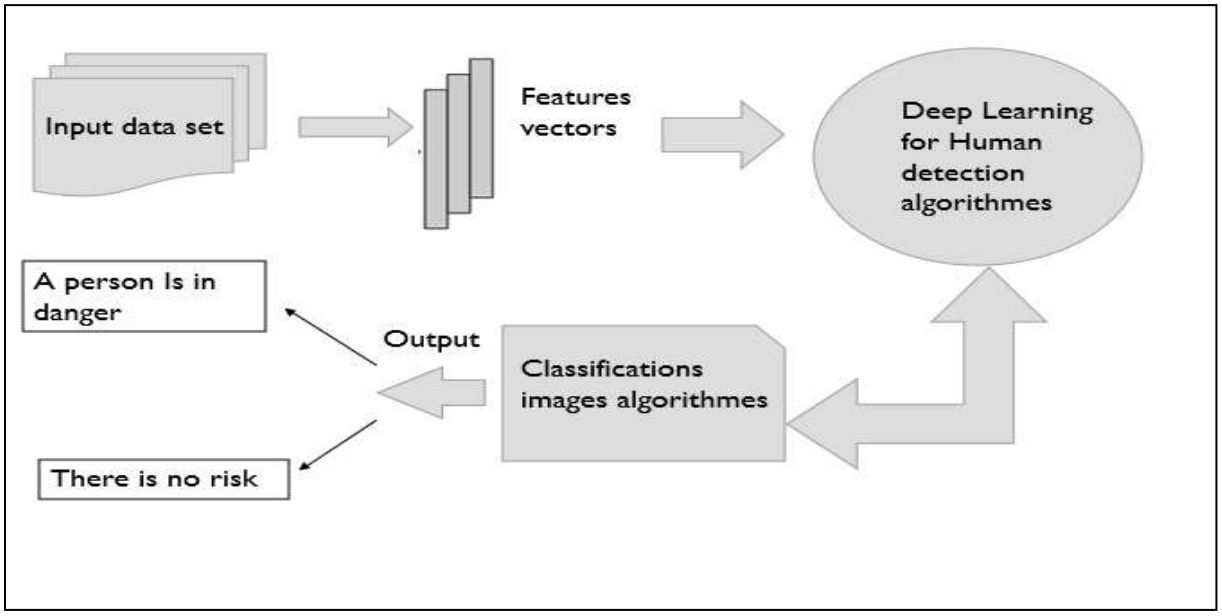


Figure 19: the main of architecture of identify the danger threatening humans

First, we will detect the person in a picture, and then we will classify his condition whether he is in danger or not.

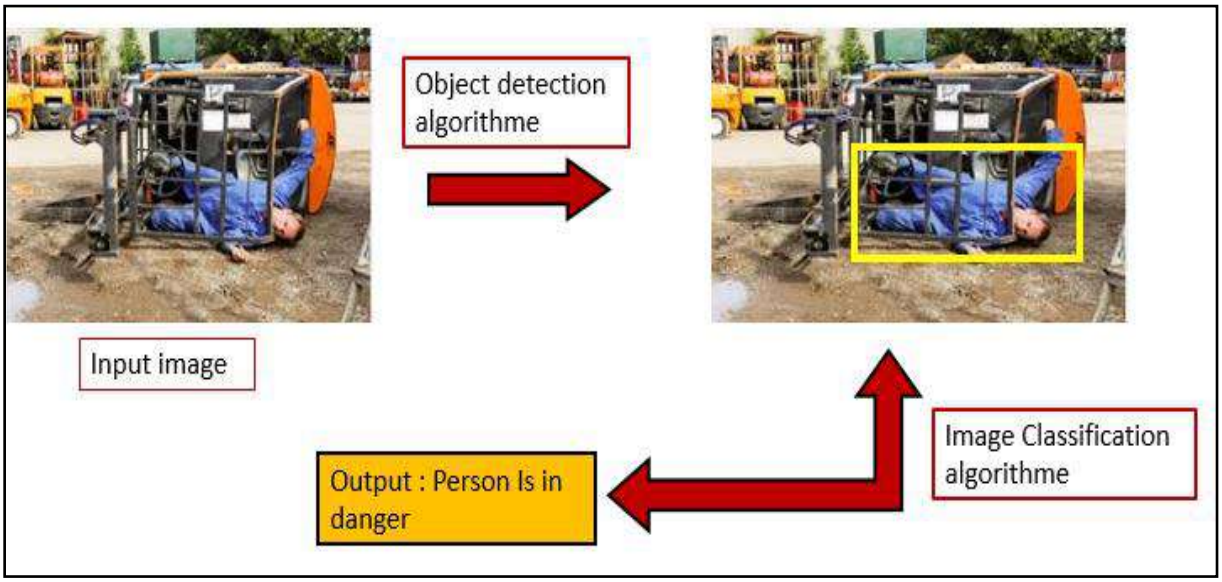


Figure 20.1: The expected result of solving our problem in the first classe

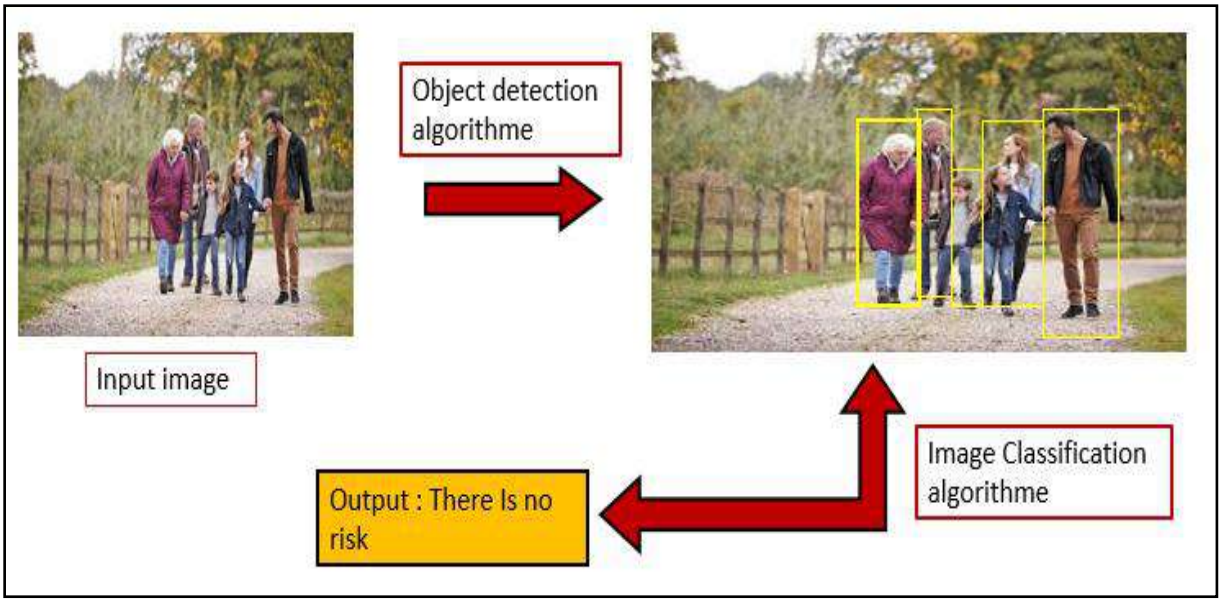


Figure 20.2: The expected result of solving our problem in the second classe

We use for detect a human in our dataset the model **YOLOV3**, and we noticed that its accuracy exceeded 99 % in detecting a human in images and from it we can rely on it in our methodology and we do not need for another model because the result is excellent .

In addition to this , we chose KNN , Vgg16 and InceptionV3 to conduct training on classifying our dataset.

3. The used dataset :

We created the new database contain tow classes, which have a relation with the recognize the danger through the smart surveillance cameras.

Our new database hold about 527 images (127 images from the first classe (person is on danger and the rest from the second classe) , this number of images is considered rather small is not enough to get an excellent result , but is the largest number that we were able to collect at the moment .

For this test set, we use Hp computer with CPU of Intel core i3 and 4GB RAM.

4. Development Tools :

Our methods were implemented using python programming language based on keras and tensorflow libraries.

4.1.Python:

Python is an interpreted, high-level and general-purpose programming language. Created by Guido van Rossum and first released in 1991, Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.[34]

We evaluated our methods using average precision (AP).

4.2.Average precision (AP) :

The mAP is a popular metric in measuring and evaluating the accuracy of object detectors algorithms like Faster R-CNN, SSD and YOLO. The mAP metric is the product of precision and recall of the detected bounding boxes. The mAP value ranges from 0 to 100. The higher number, the better it is. The mAP can be computed by calculating average precision (AP) separately for each class, then the average over the classes. A detection is considered a true positive only if the mAP is above 0.5. All detections from the test images can be combined by drawing a draw precision/recall curve for each class. The final area under the curve can be used for the comparison of algorithms. The mAP is a good measure of the sensitivity of the network while not raising many false alarms. [41]

Precision: measures how accurate are the predictions, i.e. the percentage of the predictions are correct. [7]

Recall: measures how good you find all the positives. [42]

5. Comparisons :

We compared our result of method Vgg16 and inceptionv3 the obtained results are presented in Table 4.

Table 6: Comparative results for vgg16, Inceptionv3 and KNN

	Methods for classification		
	Lazy Learners	**Eager Learners**	
	KNN	Vgg16	InceptionV3
Time trained		161.23 s	355.33 s
Accuracy	34.09%	75.18 %	93.65 %
loss		53.78%	43.92 %

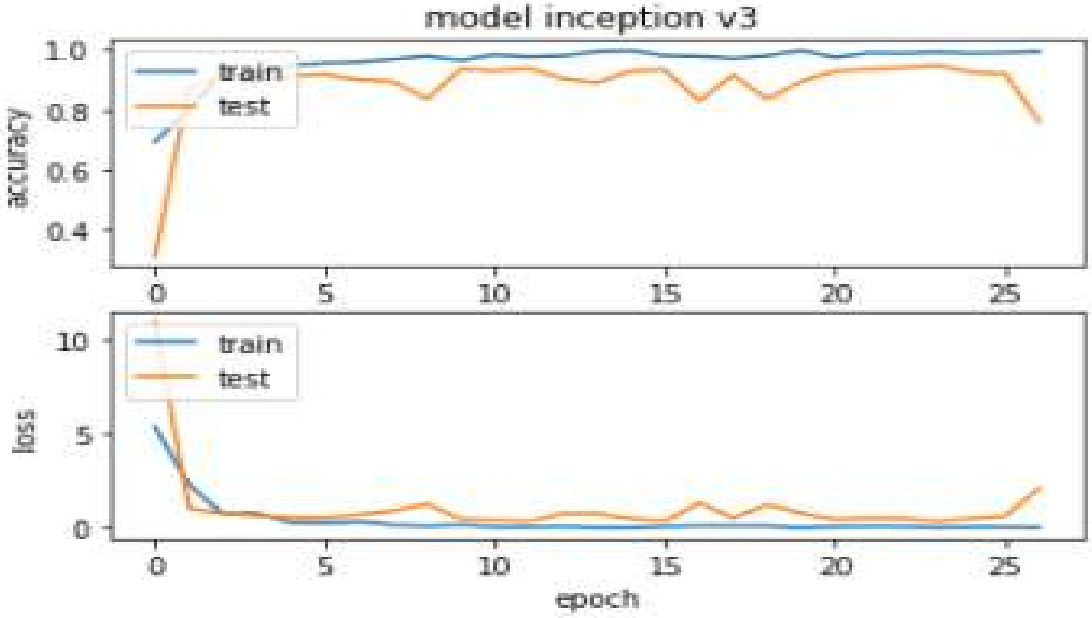


Figure 21.1: comparative result of Inception V3 model

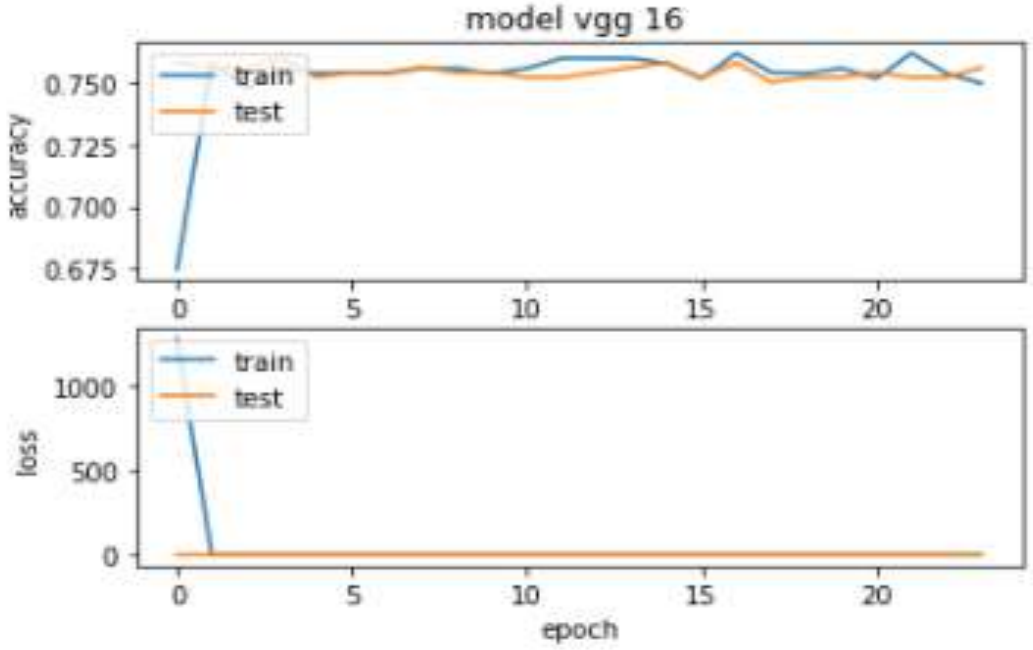


Figure 21.2: comparative result of vgg16 model

6. Discussion :

For the obtained results, the dataset arrangement that we used for the evaluation is 527 images.

The methodology that we have proposed contributes in some way to collecting some information for a good semantic interpretation of the content of images in order to preserve human lives and for a quick intervention in the event that a person experiences a work accident for example, we found that the accuracy of deep learning models (vgg16 , inceptionv3) that we used for solve this problem was better as compared to simple machine learning algorithms (KNN) ,however ,these results we obtained are not sufficient, and this indicates that the model needs more. From training time and data in those categories so, that it can provide a good semantic interpretation for this type of images.

As we expected that the accuracy could be increased more, the obtained results were promising. Hence, the proposed method classifies a person whether he is in danger or not which is acceptable.

7. Conclusion :

In this chapter we presented how works our methodology and the experiment results obtained and all the used materials, the dataset that we used, the evaluation metrics and finally we presented a comparison of the results obtained

Chapter 5 :

General conclusion

Object detection is important in the field of interpretation and semantic indexing of image content due to its many applications such as self-driving, recognition of human behavior and smart farming...Etc. Currently, no optimal solution to the problem of automatique security for camera serviance because no universally accepted is yet known approach to map low-level feature into high-level image semantic interpretation.

Despite great efforts has been put in this area achieving a images semantic interpretation rate comparable to human performance remains far away. For this reason, this area of research is still a fertile ground for future work.

This thesis was devoted to the study and realization of an automatique security system, we focused basically on developing an approach for detecting human objects in images based on **YOLOV3**. After that developing image classification for two classes (the first one human in a danger and the second no danger). Our tests were carried out on the our new dataset for humans work accidents.

We proposed a combination methods of image classification and object detection for obtain a good result.

But before this suggesting , we reviewed the difference between image classification and object detection , and the features descriptors like HOG descriptor and how it extract features ,We also explained machine learning in detail his types and his methods (KNN , SVM , K-means)and how they work also the deep learning based methods that have been used for object detection which consist in R-CNN, Fast R-CNN, Faster R-CNN, , YOLO, SSD, Mask R-CNN as well presented briefly how each technique works. And also the deep learning based methods that have been used for image classification which consist in Vgg16 and InceptionV3

Finally, we reviewed the obtained results previously in human detection and all the used materials starting by the used dataset which PASCAL VOC and COCO 2017.

The obtained results were acceptable, our methods achieved accuracy 93.65% with Inceptionv3 and vgg16 75.18%, KNN 34.09% for classification to our dataset.

Finally, Image interpretation contains problems that are not clear till now for obtain a program that identifies a person at risk or has had a work accident with high accuracy, that can be solved future work.

Perspective:

From the obtained results, we have noticed that several ideas in this field can be realized in order to increase the accuracy, which consist in many changes we are looking to make are:

- Create a new model for Image classification real time.
- Increase detection of other objects.
- More data.
- More classes.

References:

[1] R. Datta, D. Joshi, and J. Li, “Image Retrieval: Ideas, Influences, and Trends of the New Age”, ACM Trans. on Computing Surveys 2008; 20.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349–1380, 2000.

[3] Abdelkader Hamadi. Utilisation du contexte pour l’indexation sémantique des images et vidéos. Intelligence artificielle [cs.AI]. Université de Grenoble, 2014. Français.

[4] Computer vision https://en.wikipedia.org/wiki/Computer_vision#Applications . Seen in 10/ 01/2020

[5] Empirical Validation towards Improved Semantic Indexing based Text Classification using Deep Learning Neeti Sangwana , Vishal Bhatnagar https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329996 Date Written: 2018

[6]what is the semantic gap in computer vision , https://en.wikipedia.org/wiki/Semantic_gap , seen in 11/01/2020

[7] Khaoula DRID, Oum kalthoum ZOUAOUI, “Object detection and recognition from images”, Kasdi Merbah Ouargla Date Written: 2019

[8] Dasiopoulou, Stamatia, et al. "Knowledge-assisted semantic video object detection." IEEE Transactions on Circuits and Systems for Video Technology 15.10 (2005): 1210–1224.

[9] Lampert, C.Blaschko, M.Hofmann, T. (2009). Efficient subwindow search: A branch and bound framework for object localization. IEEE Trans- actions on Pattern Analysis and Machine Intelligence 31(12) 2129-2142

[10] Image classification, https://www.sciencedirect.com/topics/computer-science/image-classification , seen in 3/03/2020

[11]Feature extraction using Histogram Oriented gradients, https://www.analyticsvidhya.com/blog/2019/09/feature-engineering-images-introduction-hog-feature-descriptor/ , seen in 14/03/2020

[12] Lowe, David G. (1999). *"Object recognition from local scale-invariant features" (PDF)*. Proceedings of the International Conference on Computer Vision. **2**. pp. 1150–1157. *Doi*:*10.1109/ICCV.1999.790410*.

[13] Support vector machine, https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ , seen 17/03/2020

[14] K nearest neighbors, https://mrmint.fr/introduction-k-nearest-neighbors , seen in 17/03/2020

[15]Support vector machine step by step, https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm , seen in 18/03/2020

[16]K_means machine learning, https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf , seen in 18/03/2020

[17] K_means clustering steps, https://www.edureka.co/blog/k-means-clustering/ , seen in 19/03/2020

[18] X. Wang 1, S. Hosseinyalamdary 21. HUMAN DETECTION BASED ON A SEQUENCE OF THERMAL IMAGES USING DEEP LEARNING **(PDF).** The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-2/W13, 2019 ISPRS Geospatial Week 2019, 10–14 June 2019, Enschede, The Netherlands

[19] How deep learning works https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/ , seen in 3/08/2020

[20] deep learning , https://www.tutorialandexample.com/deep-learning-tutorial/ , seen 3/08/2020

[21] Single-Stage and Two-Stage Object Detectors ,

https://arxiv.org/abs/1803.08707 , seen in 5/08/2020

[22] Review of Deep Learning Algorithms for Object Detection,https://medium.com/comet-app/review-of-deep-learning-algorithms-for-object detection-c1f3d437b852 , seen in 8/08/2020

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014.

[24] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers2, and A.W.M. Smeulders2" Selective Search for Object Recognition" in http://www.huppelen.nl/publications/selectiveSearchDraft.pdf

[25] Ross Girshick, Jeff Donahue, Student Member, IEEE, Trevor Darrell, Member, IEEE, and Jitendra Malik, Fellow, IEEE "Region-based Convolutional Networks for Accurate Object Detection and Segmentation" in http://islab.ulsan.ac.kr/files/announcement/513/rcnn_pami.pdf

[26]Deep Learning for Object Detection: AComprehensiveReview,https://towardsdatascience.com/deep-learning-for-object-detection-a comprehensive-review-73930816d8d9, seen in 10/08/2020

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014.

[28] Object detection, https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/, seen in 10/08/2020

[29] Wei Liu , Dragomir Anguelov , Dumitru Erhan , Christian Szegedy, Scott Reed, Cheng-Yang Fu , Alexander C. Berg "SSD: Single Shot MultiBox Detector" arXiv:1512.02325v5 [cs.CV] 29 Dec 2016 in https://arxiv.org/pdf/1512.02325.pdf

[30] Review of Deep Learning Algorithms for Object Detection,https://medium.com/comet-app/review-of-deep-learning-algorithms-for-object detection-c1f3d437b852, seen in 10/08/2020

[31]Deep Learning for Object Detection: A Comprehensive Review,https://towardsdatascience.com/deep-learning-for-object-detection-a comprehensive-review-73930816d8d9, seen in 11/08/2020

[32] Zhong-Qiu Zhao, Member, IEEE, Peng Zheng, Shou-tao Xu, and Xindong Wu, Fellow, IEEE "Object Detection with Deep Learning: A Review", arXiv: 1807.05511v1 [cs.CV].

[33] yolo algorithme https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper. Seen in 12/08/2020

[34] Kuhlman, *Dave.* "A Python Book : Beginning Python, Advanced Python, and Python Exercises". *Section 1.1. Archived from* the original (PDF)

[38] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2007 (voc 2007) results (2007)," 2008.

[39] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2012 (voc2012) results (2012)," in http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html, 2011.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in ECCV, 2014

[41] mAP (mean Average Precision) for Object Detection, https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173

[42] The mean average precision Deep Learning for Computer Vision book by Rajalingappaa Shanmugamani,https://www.oreilly.com/library/view/deep-learning-for/9781788295628/089aeeb5-7e54-42dd-8c67-8853b937b1f8.xhtml,

[43] M Manoj krishna1*, M Neelima2, M Harshali3, M Venu Gopala Rao4, Image classification using Deep learning (pdf)

[44] Model vgg16, https://www.geeksforgeeks.org/vgg-16-cnn-model/ , seen in 10/09/2020

[45] Model inceptionv3, https://paperswithcode.com/method/inception-v3 , seen in 11/09/2020