UNIVERSITÀ DEGLI STUDI DI TRENTO

Department of Information Engineering
and Computer Science

Master's Degree in

Computer Science and Technologies

FINAL DISSERTATION

Integration of multiple deep learning algorithms for real-time tracking of a person in complex scenarios

Professor
Luigi Palopoli

Student
Stefano Leonardi

Co-Supervisor
Stefano Divan

Co-Supervisor
Fabiano Zenatti

Academic year 2019/2020

Ringraziamenti

...thanks to...

Contents

List of Figures **3**

List of Tables **4**

List of Acronyms **5**

1 Sommario **7**

2 Introduction **8**
2.1 Physical context 8
2.2 The Problem 8
2.2.1 Robot only environment 8
2.2.2 Environment shared between robots and humans 9
2.2.3 Purpose of the internship 9
2.2.4 Technical problems 10
2.3 The Solution 10
2.3.1 Existing technologies 10
2.3.2 Limitation of known technologies 11
2.3.3 Combine known methods to solve the general task 12
2.4 Structure of the thesis 13

3 Object Detection **14**
3.1 Task definition 14
3.1.1 Similar tasks 14
3.2 State of the art algorithms 15
3.2.1 YOLO (You Only Look Once) 15
3.2.2 SSD (Single Shot multibox Detector) 18
3.2.3 R-CNN (Region-based Convolutional Neural Networks) 19
3.3 Other famous algorithms 20
3.3.1 Mask R-CNN 20
3.3.2 Open Pose 21
3.4 Overview of the algorithms 22

4 Object Tracking **25**
4.1 Task definition 25
4.1.1 Interface 25
4.1.2 Subject of the tracking 26
4.1.3 Deal with special conditions 26
4.1.4 The traditional tracking problem compared to the thesis project . . 29

4.2 Principal known algorithms . . . 31
4.2.1 MIL (Multiple Instance Learning) tracker . . . 31
4.2.2 KCF (Kernelized Correlation Filters) tracker . . . 31
4.2.3 Median Flow tracker . . . 32
4.2.4 CSRT (Channel and Spatial Reliability Tracker) . . . 33
4.2.5 MOSSE (Minimum Output Sum of Squared Error) tracker . . . 34
4.2.6 GOTURN (Generic Object Tracking Using Regression Networks) . 35
4.2.7 TLD (Tracking-Learning-Detection) . . . 35
4.3 Which tracker could be chosen . . . 36

5 People Recognition **39**
5.1 Problem definition . . . 39
5.1.1 Video surveillance application . . . 39
5.2 Not working methods . . . 40
5.2.1 Offline algorithms . . . 40
5.2.2 Key points matching . . . 40
5.3 KNN (K-Nearest Neighbors) with images into N-dimensional space . . . 43
5.3.1 Image classifiers for representative points from images . . . 44
5.3.2 Examples of KNN applied to people recognition . . . 47

6 Solution **49**

7 Conclusions **50**

Bibliography **51**

List of Figures

3.1 Object detection applied on a sample image. 14
3.2 Similar problems respect to object localization/detection. 15
3.3 The steps of the YOLO algorithm. 17
3.4 Examples of application of NMS post-processing. 17
3.5 The steps of the SSD algorithm. 18
3.6 Comparison of the architectures of SSD and YOLO. 19
3.7 The schemes of convolutions introduced by mobileNet. 19
3.8 The 2-step elaboration of R-CNN model on a sample image. 20
3.9 Instance segmentation of an image using mask R-CNN. 21
3.10 The steps of the OpenPose algorithm. 22
3.11 Examples of application of OpenPose. 23
3.12 YOLO, SSD, Mask R-CNN and OpenPose applied on the same image. . . 24

4.1 Some visual examples of the tracking challenge conditions. 30
4.2 Examples of partial occlusion tracking. 32
4.3 The bag of bounding boxes introduced as MIL novelty. 32
4.4 The key principle of the Median Flow tracker algorithm. 33
4.5 The general intuition of the CSRT algorithm. 34
4.6 An example of application of CSRT on a frame. 35
4.7 The overall procedure of the GOTURN algorithm. 36
4.8 Examples of application of the GOTURN algorithm. 37
4.9 The interconnection of the three foundation methods of TLD algorithm. . . 37
4.10 TLD algorithm applied on a total occlusion video clip. 37

5.1 Graphical representation of SVM. 40
5.2 Three possible regions for key points: flat area, edge or corner. 40
5.3 Key points localization on a Tesla. 42
5.4 Key points matching with Notre Dame comparison. 42
5.5 Key points matching samples of people comparison. 43
5.6 Example of application of the KNN classifier. 44
5.7 The residual block of ResNet. 44
5.8 The inception module presented with the GoogLeNet model. 46
5.9 Examples of saliency and semantic parsing elaborations. 46
5.10 KNN applied with image classifier to solve the person re-identification task. 48

List of Tables

3.1 Comparison of mAP and FPS for the object detector algorithms. 24

4.1 The FPS rates of the tracking algorithms. 38

List of Acronyms

- BFM (Brute Force Matcher)
- BRIEF (Binary Robust Independent Elementary Features)
- CNN (Convolutional Neural Network)
- CPU (Central Processing Unit)
- CSRT (Channel and Spatial Reliability Tracker)
- DCF-CSR (Discriminative Correlation Filter with Channel and Spatial Reliability)
- DNN (Deep Neural Network)
- FAST (Features from Accelerated Segment Test)
- FCN (Fully Convolutional Network)
- FFT (Fast Fourier Transformations)
- FLANN (Fast Library for Approximate Nearest Neighbors)
- FoV (Field of View)
- FPS (Frames Per Second)
- GoogLeNet (Google LeNetwork)
- GOTURN (Generic Object Tracking Using Regression Networks)
- GPU (Graphics Processing Unit)
- IoU (Intersection Over Union)
- KCF (Kernelized Correlation Filters)
- KNN (K-Nearest Neighbors)
- LIDAR (Laser Imaging Detection And Ranging)
- mAP (mean Average Precision)
- MIL (Multiple Instance Learning)
- MM (Motion Model)
- MOSSE (Minimum Output Sum of Squared Error)
- NMS (Non-Maximum Suppression)

- NN (Neural Network)
- ORB (Oriented FAST and Rotated BRIEF)
- PAF (Part Affinity Fields)
- PRID (Person Re-IDentification)
- R-CNN (Region-based Convolutional Neural Networks)
- ResNet (Residual Network)
- RPN (Region Proposal Network)
- SIFT (Scale-Invariant Feature Transform)
- SSD (Single Shot multibox Detector)
- SSP-ReID (Saliency-Semantic Parsing Re-IDentification)
- STAF (Spatio-Temporal Affinity Fields)
- SURF (Speeded-Up Robust Features)
- SVM (Support Vector Machine)
- TLD (Tracking-Learning-Detection)
- YOLO (You Only Look Once)

1 Sommario

Sommario è un breve riassunto del lavoro svolto dove si descrive l'obiettivo, l'oggetto della tesi, le metodologie e le tecniche usate, i dati elaborati e la spiegazione delle conclusioni alle quali siete arrivati.

Il sommario dell'elaborato consiste al massimo di 3 pagine e deve contenere le seguenti informazioni:

- contesto e motivazioni
- breve riassunto del problema affrontato
- tecniche utilizzate e/o sviluppate
- risultati raggiunti, sottolineando il contributo personale del laureando/a

2 Introduction

This chapter offers an overview of the project on which the thesis is based. The goal is to explain in detail how the practical problem has been approached in order to analyze the physical constraints, ideate a software method able to solve them, and how these ideas were then implemented into a working algorithm.

2.1 Physical context

The physical component in this project is a robot. Its definition can vary a lot basing on the context in which it is used. For this project, a robot can be described as a vehicle able to move in the space. A **LIDAR (Laser Imaging Detection And Ranging)** sensor is mounted. Which allows to drive in the space avoiding physical obstacles during the movement. In addition, it is installed a computational device connected to a webcam that can record streams of images representing the space in front of the robot itself.
The video camera becomes the eye of the robot itself, and the captured video stream is used as the input of the algorithm working on the computational device. This computer can be composed of a **CPU (Central Processing Unit)** or more likely it is built with a **GPU (Graphics Processing Unit)** that can speed up parallelized computation, applied on the **DNNs (Deep Neural Networks)** used as the core of the algorithm. The software does not assume one component over the other, the only variation is in the performances: a GPU computation speed can be much higher than a CPU.
Instead, the output of the algorithm is a position composed of X and Y pixel coordinates calculated on a single frame captured from the webcam. This location can be then elaborated and, with the use of LIDAR sensor, the robot can estimate which is the 3D position of the element tracked from the software.
Finally, it moves to reach that position, in order to follow the tracked subject not only into the virtual space but also into the real environment.

2.2 The Problem

The thesis project is based on an internship with Dolomiti Robotics[3], a company working on self-driving robots.
These vehicles are designed to work in an industrial environment. This scenario is populated not only by robots but also people, making the driving task even more complex to achieve.

2.2.1 Robot only environment

A completely automated environment, where humans cannot access looks to be a similar context. Instead, it is completely different because each vehicle has is its own logic that can be designed to fit the requirements of all the other robots working in that area.
The typical solutions to drive a vehicle in this scenario are two:

- Based on a centralized decision unit that moves all the robots simultaneously around. This unit is responsible for avoiding collisions by knowing the exact position of each

single moving robot.

- Based on fixed rules of movement that each robot has to follow. The rules do not allow collision and the automated vehicles respect them.

Both these methods work because an automated vehicle uses a deterministic decision process and does not take arbitrary choices.

2.2.2 Environment shared between robots and humans

Instead, in a shared environment, there are a lot of elements that are not controlled by a deterministic rule. The changes in the scenario are random and prediction cannot be done. There are both fixed object that may have changed position due to external interaction, and also human that walk around with no defined rules.
In this scenario, it is fundamental to choose an input method that can measure the area around, in order to create an autonomous moving vehicle. Therefore LIDAR has been chosen. LIDAR is a technology that measures distances around the robot in a horizontal plane. The effect is that the robot knows in each direction which is the distance from the surrounding objects. This key idea has been used from Dolomiti Robotics, to design a software able to drive robots around avoiding collision with fixed obstacles or people walking.
While a robot moves around, it can construct a map of the fixed objects in the environment, measured with LIDAR. Instead, the moving objects, such as other robots or people, that are recognised as not fixed elements are not stored in the map as obstacles. This reconstruction allows the vehicle to move autonomously from one position to another knowing exactly which path to follow to reach the destination.

2.2.3 Purpose of the internship

The shared environment does not offer any real human-robot exchange. The two parts only share the same spaces. The purpose of this thesis project is to create a physical interaction between the two.
The goal is to create a new functionality *"that allows a robot to follow a person into the real environment"*. How it works:

- Track/follow is the interaction of a robot and a single person (called from now on **Leader**).

- The Leader starts the *"follow"* functionality standing in front of the webcam of the robot.

- The robot has few seconds to recognise the person inside the camera **FOV (Field Of View)** as the Leader.

- Then the Leader can freely move around in the space.

- In the meanwhile the algorithm is processing the webcam stream of images recognising the position of the Leader and start tracking it in the virtual space, while following it in the real one.

- The tracking continues for a long period, up to minutes until this functionality is stopped.

2.2.4 Technical problems

This "follow" functionality may be easily solved under certain conditions. However, solving the general scenario, it is a much harder task.
Below is listed a small collection of the principal problems that make this functionality an extremely general one, therefore hard to solve.

- The tracking should be done in real-time. It is impossible to follow a person if the processing speed is too slow. A high **FPS (Frames Per Second)** rate should be respected.
- The robot needs to physically follow the person meaning that the webcam cannot be fixed. By consequence, also the background is not fixed and the entire captured image, subject included, might be blurred.
- The person can move freely around walking fast, slow, or staying.
- The Leader is a random person, it is not known while the algorithm was designed (no parameters can be fixed in advance).
- While the Leader is walking around there might be also other people that interfere with the algorithm.
- The Leader can be hidden from the webcam due to moving or static elements placed between the Leader and the webcam itself.
- The Leader can exit the field of view of the webcam disappearing until the robot rotates to watch it back again.
- The tracking should be performed for a long period.

2.3 The Solution

The problem is complex due to its generality and the necessity to cover a lot of complementary conditions. For this reason more than one solution exists. In this thesis is presented a solution based on the combination of three methods. Each one is designed to solve sub-problem compared to this one, and none of them alone can overcome the challenge of the general task.

2.3.1 Existing technologies

The three technologies are:

- **Object Detection (or Localization):** given an image the object detection task aim at processing the image and recognise which objects exist there. The detection not only need to produce a list of all the classes[1] of objects visible in the image, but also recognise in which section of the frame every single element is.
  The output of detection is a list of: class to which the element belongs, the probability associated and the **bounding box** defined as the smallest rectangle that contains the entire element.

[1] There are a set of types to which each element can be associated i.e. person, dog, car, bicycle, bottle and so on.

- **Object Tracking:** in this case the input is not a single image but a video stream and an initial section[2] of it. The goal is to remember this portion of the image and recognise it in all the frames after the first one. It is important to note that the tracking procedure it is not designed to follow a person, a car or other it is designed to follow a rectangle of coloured pixels, no matter what these pixels represent.

- **Object Recognition:** this is a comparison between several pictures. These often represent a bounding box of the object that needs to be recognised. The procedure has a database of images each one with a specific class, and the input value is another picture, called **query**, that does not exist in the database but it represents a subject known. The goal is to extract from the database all the images that have a subject that looks similar to the one represented in the query.
  This application is mainly used to recognise humans, often in the video surveillance context. The database is composed of the bounding box of all the people seen, i.e. in a supermarket over the last week, and when a thief is captured and it is used as a query. So, the system should return all the images containing the thief itself.

2.3.2 Limitation of known technologies

The challenges presented previously in Section 2.3.1, can solve a small part of the general problem but each one has a technical problem Section 2.2.4 that cannot be solved:

- Object detection is a computationally expensive task, on a powerful GPU can run in real-time but that is not the case of the robots we are working with.
  In addition, the detection works frame by frame and each one is independent of the previous one. So, if a person is recognised in a frame, and in the next one, there is more than a single person the algorithm does not know the relation between all of them. Meaning that a person cannot be tracked from one frame to the next one.

- Object tracking, according to the name, seems the task that better match the requirement of the general problem.
  Despite that, the tracking does not consider that the tracked subject, the Leader, cannot be hidden from the webcam. The Leader should always be visible into the recorded video, and that is not the case. In addition, the Leader can also exit the field of view of the robot while walking around.
  Lastly, all the trackers are designed to follow the subject for small periods[3], after a while the tracked rectangle of coloured pixel changes and the precision of the output is no more not guaranteed. This phenomenon is known as the **drift effect**, after a while the drift is so wide that the tracked cannot be trusted anymore.

- Object recognition due to its requirement was not designed at all to run in real-time. In fact, it is enough to run this procedure only when a query occurs, and that does not happen more than one every second.
  Except that, there is a more intrinsic problem with the recognition to approach the general problem. The procedure requires a query that can be the subject at the actual frame, but then it should work on a dataset composed of old frames and these are useless to solve the actual frame.
  In addition, this algorithm cannot be independent, because the input values are

[2] A portion of the image: a rectangle.
[3] Each tracker works on a video of few seconds.

bounding boxes of people, but these regions can be computed only with an object detection algorithm. Hence this approach cannot solve the problem independently.

This explanation shows that none of the existing proposed technologies can solve the general problem in all its parts.

2.3.3 Combine known methods to solve the general task

To solve the problem and manage all the requirements it is necessary to create a combination of known methods.
An example of integration of methods to solve a complex task was done by Jiang et al, in their paper[4] that presents a fusion of **YOLO9000**[5] (the second version of **YOLO**[6]) used as object detector and **SURF**[7] used as short-term object tracker.
The paper illustrate an innovative approach based on two thresholds that are used to understand when the drift of the tracker is too large and it is necessary to reinitialize it. So YOLO is executed to find the tracked subject back again and after the initialization the loop can start again.

The method presented in that paper is an integration of two class of methods. Instead in this thesis is presented an integration of three. The third method is necessary because an additional technical problem exists (Section 2.2.4). Jiang et al. work on sport video clips where athletes are always followed by the camera and never disappear out of the field of view. In addition, occlusion can exist but are very short and the tracker is often able to overcome them.
Instead, in our scenario we need to manage the disappearance of the Leader behind a corner for a relatively long time. So, the object recognition method was introduced to solve this condition.
These are the main steps of the entire algorithm:

1. The detection is executed and the bounding box of the Leader is found. By assumptions, in this phase, if more than one person is simultaneously found the detections are ignored.

2. The tracker is initialized with the bounding box found.

3. The tracker runs for the next F frames.

4. A new detection is executed and D people are found.

5. The person recognition is used to choose if the Leader is contained in the list of people found:

   - If **yes**: the procedure starts again from point 2 (tracking)
   - If **not**: the procedure loops again from point 4 (detection again)

This flow shows how detection, tracking and recognition are combined together to build a complete algorithm, that can run in real-time due to the alternation of slow and fast methods and to manage all the problematic scenarios.
The details will follow.

2.4 Structure of the thesis

The next chapters are organized as follows. This section concludes the introduction (Chapter 2).
Then follow three chapters one for each main method: object detection in Chapter 3, object tracking in Chapter 4 and object recognition in Chapter 5.
An overview of the entire algorithm and how it works together follow in Chapter 6.
In the end, the conclusions are presented in Chapter 7.

3 Object Detection

This chapter explains into details what is the object detection task and the methods that can solve it efficiently. Moreover an overview of other methods is given among whose there not efficient ones and others that may seem to solve the task but do not.

3.1 Task definition

Object detection, also known as **object localization**, is an evolution of the **image classification** (Figure 3.2). In classification, an algorithm should produce a list of all the classes of objects inside the image. Instead, the detection not only calculates which object class exists but also how many occurrences are present for each class. Then, the complex part, and the most interesting one for this thesis application, is localizing where those elements are placed inside the image. The position is not considered as a point but as a bounding box defined as the smallest rectangle that contains the entire element. An example of object detection is shown in Figure 3.1.

Figure 3.1: Object detection applied on a sample image.

3.1.1 Similar tasks

Object detection can be additionally improved to extract even more information from an image.
The main evolutions, shown in Figure 3.2 are:

- **Semantic segmentation:** takes all the bounding boxes produced by an object

detector, and for each one, it calculates the pixels that belong to the object itself and the ones that do not. By doing this each class has its own colour associated. As a result, the algorithm knows for each pixel if it belongs to one label associated with the image (semantic division) or to the background (yellow in the image).

- **Instance segmentation:** is similar to semantic segmentation, but in this case, each instance of the object is considered as a new element. In fact, the three cubes in the figure have different colours associated to them.
  This task is solved by the **Mask-R-CNN algorithm** (Section 3.3.1).

- **(Human) pose estimation:** is the most complex task among the five. Mainly applied to people, this challenge consists in the estimation of the 3D position of the body. The idea is to build up a skeleton of the person in the image and understand how its body limbs are positioned. This functionality is important to understand what a person is doing in the image.
  This task is solved by the **OpenPose estimation algorithm** (Section 3.3.2).

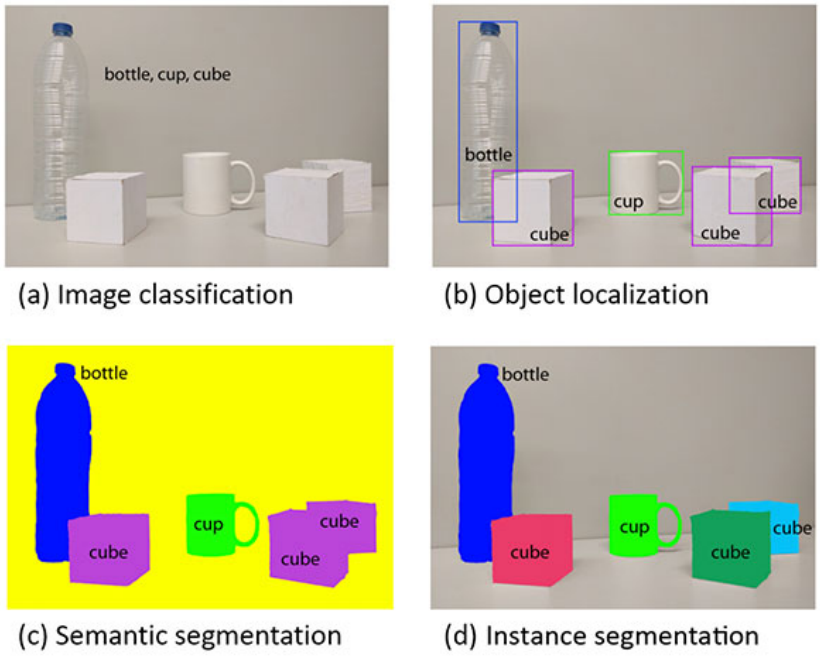


Figure 3.2: Similar problems respect to object localization/detection.

3.2 State of the art algorithms

Object detection has a lot of applications both in real-time, such as in this thesis, but also in safety-critical scenarios like cars with autonomous driving. This division brings out two different metrics: precision and speed. The ideal detector is both fast and precise, however this algorithm does not exist yet. The methods can be divided into two categories.
The solutions mainly focus on speed: YOLO (Section 3.2.1) and SSD (Section 3.2.2).
Instead, the one mainly focused on precision is R-CNN (Section 3.2.3).

3.2.1 YOLO (You Only Look Once)

YOLO[6] was initially designed in 2016. At that time it was the first object detector approach to use a single **CNN (Convolutional Neural Network)**. Redmon et al.

goals were to create an extremely fast detector. An overview of the overall procedure is shown in Figure 3.3, and the architecture appears in Figure 3.6.
The image shows a two step procedure, but these steps are solved in parallel. This is the core idea of the paper. A single CNN can be highly optimized.
The YOLO procedure works as follows[1]:

- Preprocess: the image is resized to fit the standard input dimension of the CNN.
- Left image: the picture it is divided into a grid of CxC cells.
- Top image: each cell suggests some bounding boxes centred on it, that can match elements in the background. To each box is associated a value describing the probability that it contains one of the elements of the image.
  At most one detection per box can be selected as correct. This relies on the assumption that two correct bounding boxes cannot share the centre. This is both an efficient idea but also a big limitation. Too small elements, close to each other, cannot be both detected.
- Bottom image: each cell has an associated probability regarding a label that represents the class that can be found in that cell if an element exists in it.
  I.e. the cyan cells means: "if there is something here, it will belong to class 'DOG' ".
- Right image: the two partial elaborations are merged. The most likely bounding boxes are chosen and classes are associated to them according to the value for each cell in the probability map.

That was the first YOLO version, in this thesis is used the third[8]. Mainly, the changes were about recognition of a wider set of classes and small implementing details to improve the overall precision of the algorithm.

The output of the CNN is generated extremely fast and it is accurate however has a big problem. Often if two classes have similar probabilities or the shape of the element is not perfect YOLO might propose more than one bounding box for each element. That is the case of Figure 3.4c where the truck is classified both as "truck" and as "car". The same happens to the person that has been seen twice.
To solve this, it is necessary to apply a new technique: Non-Maximum Suppression.

NMS (Non-Maximum Suppression)

This technique[9] is a post-processing that works on the bounding boxes suggested, as output, from YOLO or other detectors. NMS does not consider the source image. The goal of this procedure is to refine the bounding boxes proposed and to choose which subset of them is better to fit the final image prediction. Two examples of applications are shown in Figure 3.4.
The main flow of the algorithm is as follows:

- The input is a list of all the boxes generated for a single image. Associated to each one there is its probability.

[1]The original presentation of YOLO by Redmon at CVPR 2016 conference can be found here.

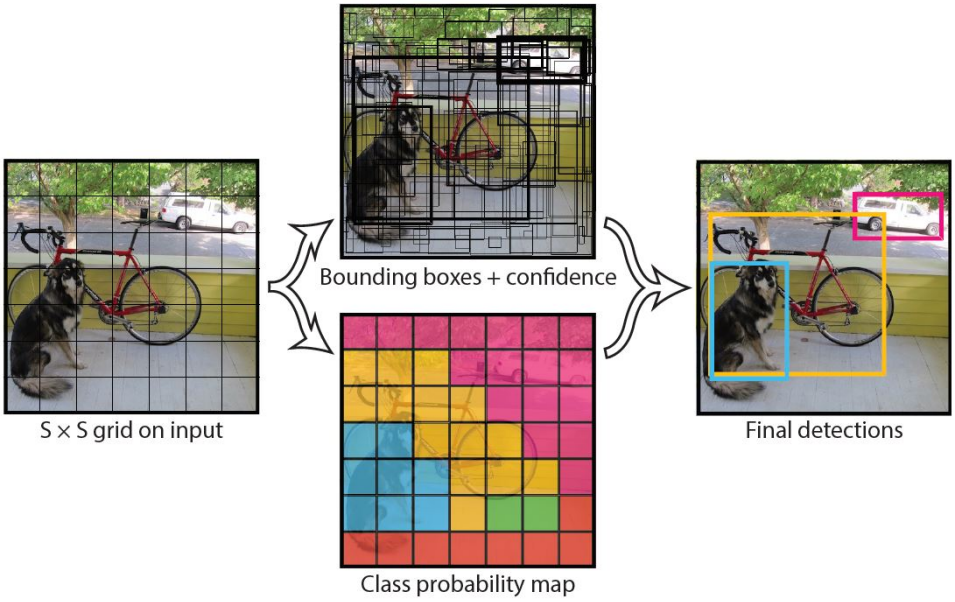
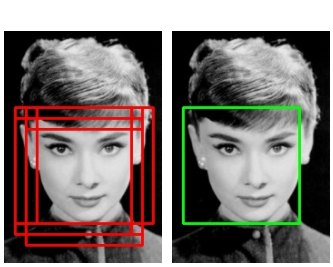


Figure 3.3: The YOLO image elaboration based on bounding box proposal and class probability map.

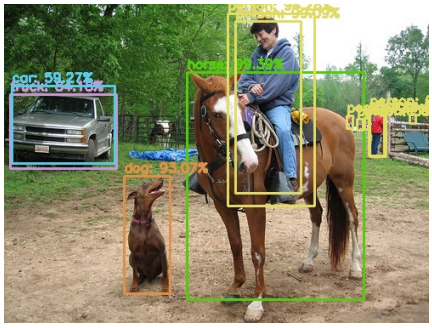
- The boxes are sorted in decreasing order according to the probability associated.
- Then, each box is accepted or rejected according to the **IoU (Intersection Over Union)**. That is the percentage of overlapping area with an already accepted box.
  - If the IoU is above a certain threshold, meaning that the two boxes overlap too much, the one with the lowest probability is discarded.
  - If that is not the case, the box is accepted as a new prediction.

The input in Figure 3.4a, is processed and only one box is accepted (Figure 3.4b) because the IoU is very high. Instead, in Figure 3.4c, two boxes are removed respectively from two other separated boxes (Figure 3.4d) because two different subjects are involved.

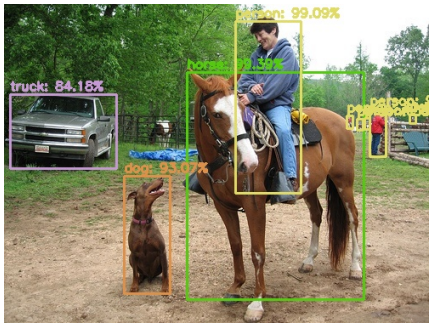


(a) Overlap bounding boxes

(b) Refined bounding box



(c) YOLO generated bounding boxes



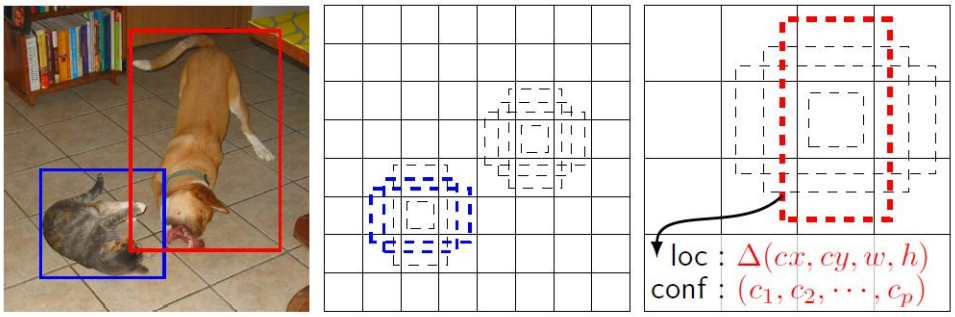
(d) Apply NMS to refine the YOLO's output

Figure 3.4: Two scenarios of application of Non-maximum suppression algorithm. First: choose which of the 6 manual generated bounding boxes, on Audrey Hepburn's face, should be considered the correct one. Second: refinement of the YOLO prediction output, by removing the "car" and "person" prediction.

3.2.2 SSD (Single Shot multibox Detector)

The principal competitor of YOLO is SSD[10]. Both are based on the same principle: the use of a single Convolutional Neural Network to propose bounding boxes and associate them to classes. The CNN is optimized as much as possible to improve the speed performance and eventually even the accuracy.
The difference relies on how the two algorithms deal with the bounding boxes suggestion. YOLO for each cell of the grid chooses a couple of options and at most one can be chosen. On the other hand, SSD works as follows (Figure 3.5):

- The image is divided into a grid of CxC cells, called **feature map**.
- Each cell can propose a set of default boxes that has a size measured in cells (i.e. 3 cells high and 2 wide).
- The process is repeated many times varying the value of C: the **granularity of the grid**. This guarantees that the algorithm is scale-independent matching both, big and small, subjects.
  In Figure 3.6 is shown how the convolution layers blocks are matched together only at the end.
- All the suggestions are merged together to produce the final proposals.
- SSD internally performs NMS to remove unnecessary detections.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Figure 3.5: The SSD image processing and how the bounding box proposal is elaborated.

MobileNet

The implementation of the project does not use a traditional version of SSD, but a lighter one. This model is a combination of SSD and mobileNet[11].
MobileNet is a methodology that approaches Convolutional Neural Networks to transform the architecture structure to build a much lighter version of the model. The concept was first ideated to allow low power devices, such as smartphones, to run computational expensive algorithms based on CNN.
The principle is to replace each standard convolution (Figure 3.7a) with a **Depthwise separable filter**. A standard convolution works on a grid of DxD pixels and for each one produces output features of depth M. This operation can be repeated N times for each source feature.
The operation is broken into two other simpler convolutions:

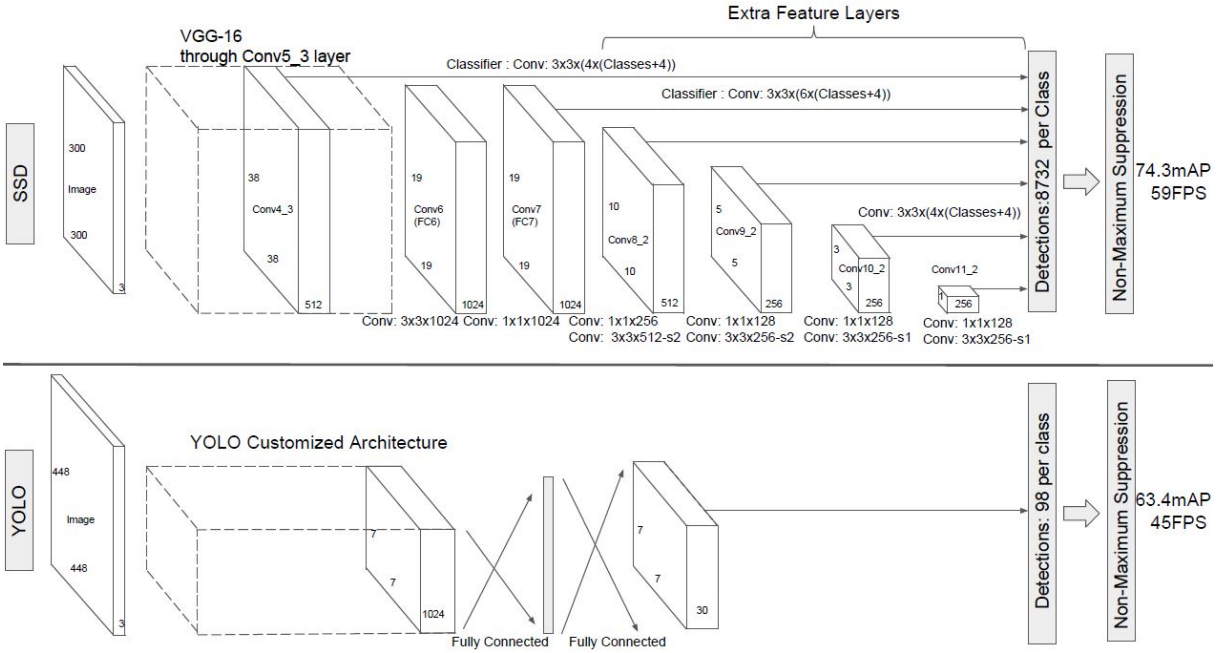
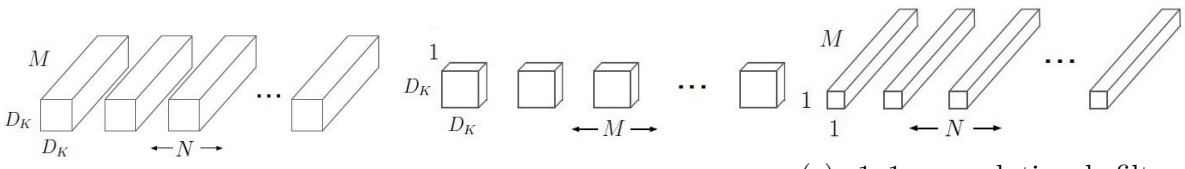


Figure 3.6: A comparison of architectures between SSD and YOLO which is designed as a compact block. Instead, SSD is modular, it is divided into convolution layers of different scales, combined at the end, to make the algorithm scale-independent.

- **Depthwise convolutional filters** (Figure 3.7b): produces only one feature output at a time, repeated M times for each DxD grid.
- **Pointwise convolution filters** (Figure 3.7c): extends the output feature of the depthwise filter to N output features.

The original paper demonstrates how these two operations stacked in a row, can produce results close to the correct ones.
The computational cost determined by the number of parameters, used by depthwise and pointwise filters, can be further reduced by randomly removing a percentage of these parameters. According to the portion of parameters removed (25%, 50%, 75%), the algorithm precision is affected.



(a) Standard convolution filters

(b) Depthwise convolutional filters

(c) 1x1 convolutional filters called Pointwise convolution filters

Figure 3.7: The novelty of mobileNet is that it converts a traditional convolution (A), into a combination of two lighter convolutions (B-C), that produce almost the same output.

3.2.3 R-CNN (Region-based Convolutional Neural Networks)

R-CNN[12] was the first invented method of the three in this section. Differently from YOLO and SSD, it is mainly focused on performing detections with high precision, despite

the processing time.
This algorithm is a two step object detector. The workflow, shown in Figure 3.8, works as follows:

1. A region proposal algorithm is executed on the input image and it produces 2000 bounding boxes.

2. Each of these proposals are elaborated independently.
   For each box, an image classifier, based on CNN, produces features from the image and then predicts which classes they might contain.
   Any kind of image classifier can be used for this task resulting in an algorithm that can be easily adapted with new networks.

The main problem is that overlapping proposals are elaborated independently. The result is that feature extraction is performed on the same area of the image multiple times. These have been solved with a second version of the algorithm, called **Fast-R-CNN**[13]. The feature extraction for the full image is performed before the image classification that now works on an already generated image of features.
An incremental improvement, comes from the third version: **Faster-R-CNN**[14]. It performs the feature extraction as the first step and based on it the bounding boxes are proposed with **RPN (Region Proposal Network)**. The modified structure allows ad hoc optimizations to improve the low processing time.

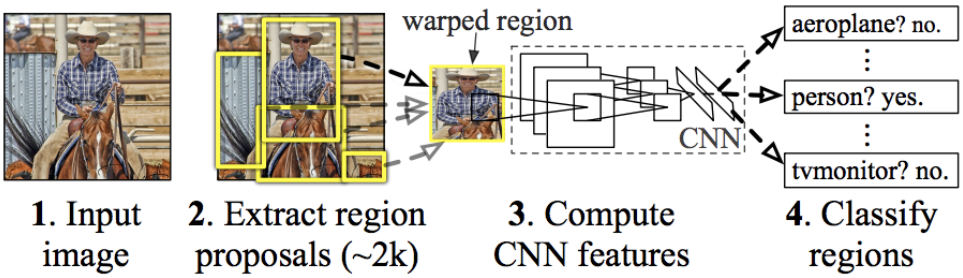


Figure 3.8: The 2-step elaboration of R-CNN model on a sample image.

3.3 Other famous algorithms

3.3.1 Mask R-CNN

A variation of R-CNN that aims to solve the instance segmentation task (Figure 3.2) is Mask R-CNN[15].
Mask R-CNN is built on top of two technologies:

- Faster R-CNN used as an object detector.

- **FCN (Fully Convolutional Network)**[16] that performs semantic segmentation.

For each bounding box, it is known the class, then FCN computes the segmentation of that class. All the shapes of elements in the image are then merged together to build the instance segmentation result. An application of this algorithm is shown in Figure 3.9.

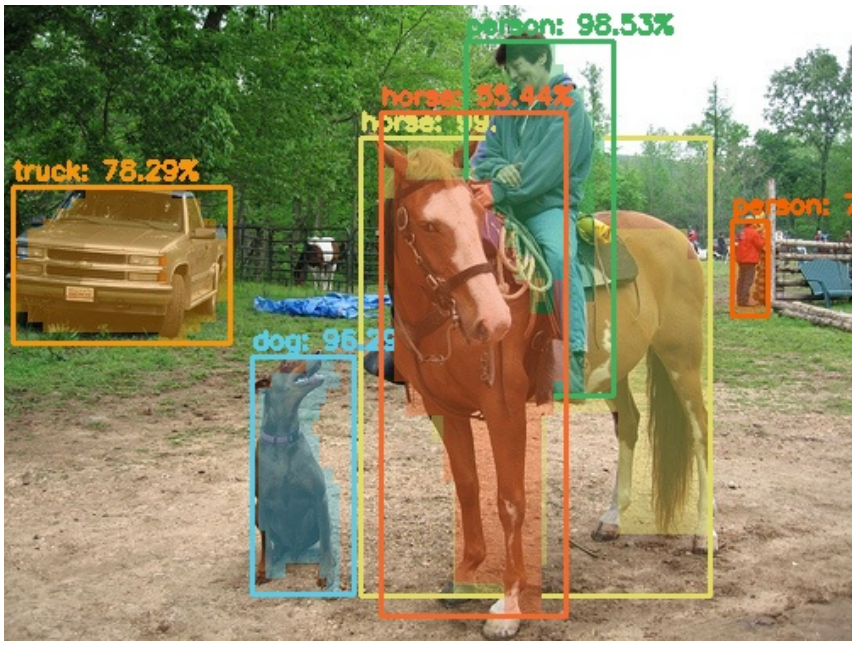


Figure 3.9: Instance segmentation of an image using mask R-CNN.

3.3.2 Open Pose

Firstly designed in 2017 OpenPose[17] aims at processing an image and recognising the position of the people in it. The position is the skeleton of a person, it is the interconnection of limbs that link 15 points on the human body.
This algorithm opens al lot of possibilities because until then estimating the body position was achieved with 3D cameras, extremely expensive hardware that now can be easily substituted. Recently, OpenPose has been improved in terms of speed, to process frames in a video with **STAF (Spatio-Temporal Affinity Field)**[18], and it is also been extended to understand the 3D human position.
An overview of the overall procedure of the algorithm is shown in Figure 3.10, instead some output examples are shown in Figure 3.11.
The OpenPose procedure works as follows[2]:

a) Preprocessing: the input image is reshaped to match the requirements of the two-branch CNN.

b) Part Confidence Maps: the first branch of the CNN process the image to extract the location of the 15 body parts.
Each body part (i.e. left shoulder, left knee, right wrist...) is detected by an ad hoc filter. These filters do not process one person at a time but the entire image simultaneously. The result is that a filter recognises all the visible right elbows in the image. This is extremely important because, by doing this, the algorithm process speed is independent respect to the number of people in the image.

c) Part affinity fields: the second branch of the CNN process the image to recognise the limbs that can connect all the body points found so far.

d) Bipartite matching: has the goal to match all the elements found to reconstruct the skeleton of the entire person.

[2]The original demo of OpenPose by Zhe Cao at CVPR 2017 conference can be found here.

This reconstruction is a greedy approach mainly based on geometry that wants to minimize the distance of two parts that must be connected.

e) Parsing results: the output image is assembled with the full-body poses for all people in the image.

It is important to note that OpenPose is a bottom-up approach. The algorithm has no knowledge about the positioning of people in the image, it only tries to reconstruct the skeleton from small sections of the body.

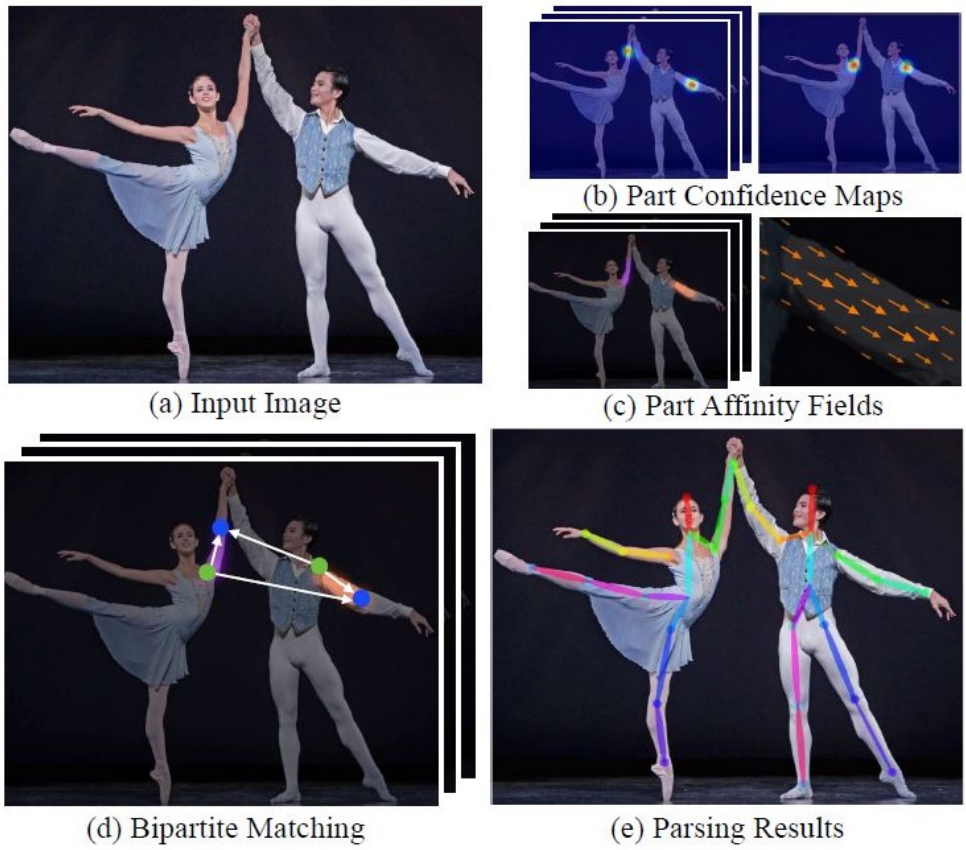


Figure 3.10: The elaboration of the OpenPose algorithm to recognise the skeleton of the two dancers in the image.

3.4 Overview of the algorithms

The algorithms presented in this chapter represent the state of the art methods for their field of application.
All those methods can be used to achieve the long-term tracking that is the goal of this thesis, but the different information generated should be used to solve the problem with different approaches. We have chosen to use the methods that perform only the detection task. This results on one hand into general information as output, and on the other hand into a very high processing speed that is fundamental for real-time application.

Performances comparison

A comparison of the speed, measured as **FPS (Frames Per Second)**, and of the precision, measured as **mAP (mean Average Precision)**, is shown in Table 3.1. The data

Figure 3.11: Some examples, taken from the original paper, on how OpenPose works.

are based on the Pascal VOC 2007 dataset[19] and come from multiple papers[6][5].
The measures do not show YOLOv3 because compared to the other methods it is more recent and a fair comparison does not exist. Instead, for Mask R-CNN and OpenPose, only the frame rate is shown because the mAP can be computed. However it is completely irrelevant respect to the other presented in the table. These two algorithms perform different tasks hence the precision of the result cannot be compared.
By looking at the data, it clearly appears that the single-stage algorithms (SSD and YOLO) are much faster respect to the two-stage methods (R-CNN), in fact, they run around 5-10 times faster than R-CNN. Instead, the precision of the three detectors is almost the same. For these reasons, we have chosen to use in this project the last version of YOLO and a lighter version of SSD: mobileNet-SSD. This idea pays a few percentage points in term of mAP but implements the CNN with fewer parameters, results in a low power consumption method. This aspect is important because the robots do not mount top quality hardware, therefore the light version of SSD can be executed more easily.

Output visualization

To visually show the potentialities of these algorithms we have applied all of them on the same picture. This elaboration is presented in Figure 3.12. Independently from the task that they solve, it appears evident that all of them recognise the 5 people in the foreground, but there are some differences:

- SSD has troubles with the player on the left. In fact, the percentage associated with him is only a 32%.
- YOLO is able to detect even a sixth person in the background that none of the others have seen.
- YOLO is trained to recognise a wide variety of objects respect to the other detectors, in fact, it is able to recognise even the "sports ball".
  Even SSD is adaptable and can be trained to recognise the ball. However the big advantage of the second version of YOLO (also called **YOLO9000**), is that it was integrated with the **Wordnet graph**[20] to be scalable in terms of the number of

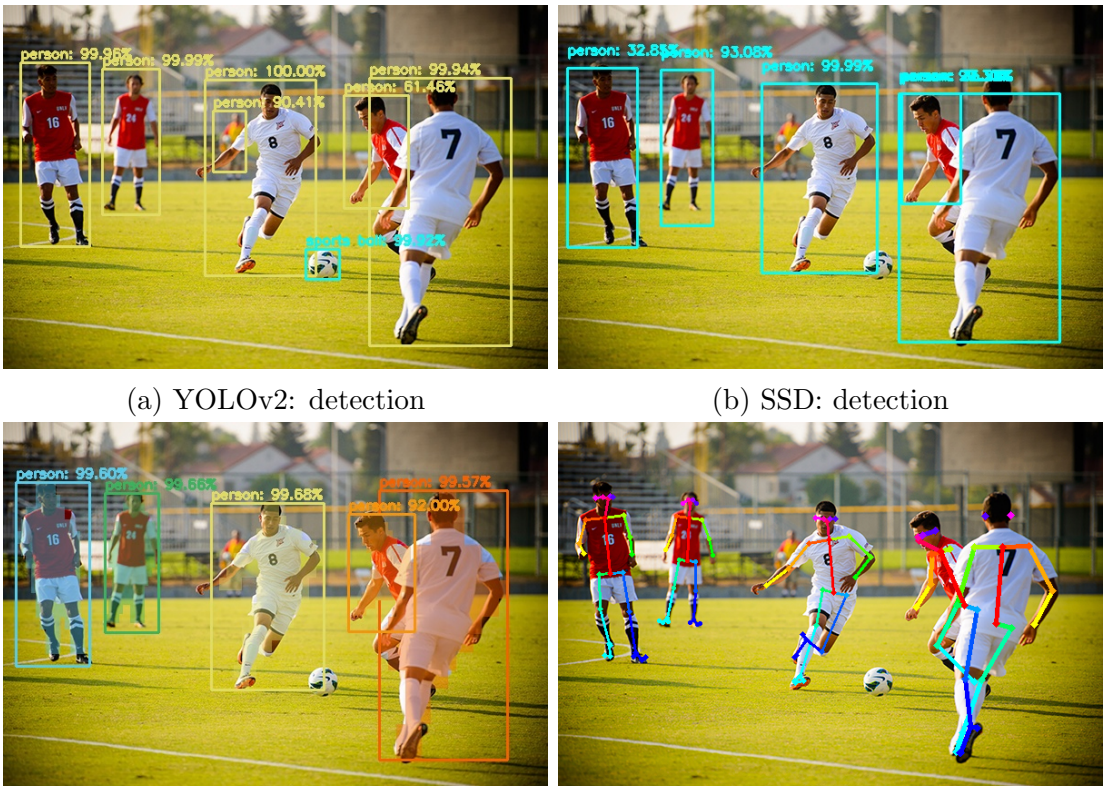
classes recognised. The result is a detector that is able to recognise up to 9000 different classes.

- Mask R-CNN is, in this example, the most accurate algorithm. It recognises four people with a precision of 99% and the last one with 92%.
- OpenPose, by estimating the body parts, can even understand the orientation of the people in the soccer field. This big advantage can be used to understand where these people will move in the frame after this one.

Considering these reasons, it is evident that discarding OpenPose and Mask R-CNN, in favour of SSD and YOLO, is only a choice for this project. All these algorithms have potentiality that can be used to create a good object tracker.

Pascal VOC 2007	YOLO			R-CNN				
Algorithm	v1	v2	SSD	R-CNN	Fast	Faster	Mask	OpenPose
FPS	45	**67**	46	0.05	0.5	7	7	10
mAP	66	**<u>76</u>**	<u>74</u>	53	70	<u>73</u>	X	X

Table 3.1: Overview of the performances of the algorithms presented. The evaluation is based on the mAP and the processing speed. In bold the best scores and underlined the mAP of the three states of the art detectors.



(a) YOLOv2: detection

(b) SSD: detection

(c) Mask R-CNN: instance segmentation

(d) OpenPose: human pose estimation

Figure 3.12: An example of application of the four main algorithms, based on the same sample image.

4 Object Tracking

This chapter is focused on a few methods that solve the traditional task of object tracking. For each algorithm the potentiality is shown and the advantages, compared to other methods, will be discussed. In addition, an overview of similar problems is given in Section 4.1.3.

4.1 Task definition

Object tracking is a challenge that, differently from the detection, does not work with images but deal with videos. A video is technically defined as a sequence of images, called **frames**, combined together at a certain frame rate. Typically this rate is 15, 30 or 60 FPS (Frames Per Second).
The interconnection of the frames of videos is used from the object tracker algorithms to create knowledge, based on the previously analysed images, by storing information. An object detector elaborates a picture at a time, each one independently from the other. There is no connection among images of the same database. Instead, a tracker process images one at a time but it uses the knowledge, of the previously analysed frames, to understand the current incoming image.
The task baseline is defined as:

> ***The traditional tracking challenge*** *consists of tracking a single well-defined subject over a short video clip, no matter if it is a person, an animal or an inanimate object. The target of the tracking clearly fully appears into all the frames of the sequence.*

4.1.1 Interface

Structure of the tracking algorithms interface:

- One input value is composed of an image, typically the first frame of a video. In our case of real-time elaboration the image corresponds to the actual view of a webcam connected to the device.

- The other input corresponds to the bounding box coordinates of the subject that needs to be tracked. The bounding box can be manually generated or can be automatically identified with an object detector.

- All future steps get as input a new frame. The task is to understand, in this new image, where is located the subject defined by the initial bounding box.
  Then, if all these elaborated images and the generated bounding boxes, are combined together, the result is a video. The initial subject of the video while moving is always centred into a rectangle that follows it across the entire **FoV (Field of View)** of the camera.

4.1.2 Subject of the tracking

It is important to note that the task is called "*Object tracking*" and not "*Person tracking*" this is because there are no limitations to the subject that needs to be tracked, it does not necessarily need to be a human. A better name for the challenge could be "*Area tracking*" because the algorithms should follow the rectangle of coloured pixels that was located in the initial bounding box. No assumption can be done of which type of subject exists inside the bounding box.
Generally, these methods work if the initial position is centred around an object that moves consistently in the space and does not change aspect. Some examples that could break these methods are:

- The initial bounding box contains two or more objects that move independently from each other. The algorithm will recognise one of the two as the main subject and will track it over the video and lose the other one.
  This is a problem if accidentally the bounding box contains something that is "more important" than the "*real subject*". It can also be an advantage that allows to easily discard the background that is an "*object*" itself.
  Since two independent objects move differently in the space compared to a single unique object, tracking them can easily create unexpected errors due to inconsistent movements.

- The subject changes aspect too rapidly due to different luminosity in the environment, or elaborated video sequences that are not "*natural*". This last possibility often breaks methods designed for real situations, so it is not a problem only in this case. Instead, the change of luminosity often occurs but a solution can be achieved with frame pre-processing, such as the standardization of the luminosity of the image to always process a figure with the same luminosity.

4.1.3 Deal with special conditions

The tracking task can be seen as a set of problems. This is because there are a lot of conditions that can modify the scenario where the algorithms should work. By modifying the type of difficulties in the videos some methods may fail while others may not.
The goal of this thesis is to build a "person tracker", that should work under certain conditions. The problem is that the combination of a lot of requirements makes the problem harder.
Despite the time spent for a programmer to ideate, implement and test a solution there is a trade-off to consider. Solving a hard task requires a more computational complex solution compared to solving a similar easier task. This complexity influences the performances of the proposed algorithm. To conclude, it is important to understand which are the requirements of the problem that we are dealing with, in order to choose the algorithm that it is better to use, to solve the task and to perform it fast.

Below are listed a set of requirements that make the baseline harder (definition is in Section 4.1):

- **Changes of 2D shape**: the target due to movements might change its ratio, aspect, and shape. We are interested in what the camera sees of the subject (2D space) and not the effective actual condition of the subject (3D space).

- Partial occlusion (Figure 4.1f): it may happen that during the video, part of the subject is occluded, by an object. If this happens the algorithms must be designed to be robust and to localize the target even with a small section of it. The bounding box generated may contain only the visible part of the subject but it may also happen that an estimation of the entire subject is done and the bounding box contains both the visible and the estimated missing area of the target.
- Rotations and deformations (Figure 4.1b): the object tracked while moving can rotate and show to the camera a different side, or if it is deformable, change the shape (i.e. a person walking change the shape continuously). This might influence even colours.

- **Total occlusion** (Figure 4.1i): the subject completely disappears behind an object or out of the camera field of view.

  - Short-time occlusion: the subject is hidden for a very small number of frames (i.e. 5 or less). This often happens when the subject or an obstacle is moving and the three elements, subject, object and camera, are aligned. These very short occlusions may be solved using a little memory that stores the subject information for the last few frames (i.e. 10).
  - Long-time occlusion: the occlusion lasts for a bigger number of frames, even seconds. Often it occurs when the subject exits the field of view of the camera and does not enter it again for a while. It can also happen if the subject is behind a big obstacle such as a vehicle or a wall.
    It is harder to solve compared to the previous scenario, because the solution requires a long term memory associated with a recognition procedure to understand when the subject is visible back again.

- **Fast-moving object**: the subject shifts for a big portion of the picture from the old position to the new one in a single frame. Meaning that there was a big 2D movement. This could happen because the subject is effectively moving fast, or because it is close to the camera and even a small movement looks big.

  - Blurred subject (Figure 4.1d): due to its movement the subject is blurred and this heavily changes its aspect. The modified elements are the shape and the colours that are somehow faded. In addition, especially for humans, moving fast can make parts of the body such as arms or legs disappear.
  - Proximity assumption (Figure 4.1g): a lot of trackers are based on the assumption that the subject moves around only a little bit. If it moves fast this principle is broken.
    After knowing the exact location of the subject on the previous frame starts the estimation on the following one.
    At this point there are two things to focus on:

    ➤ If the subject does not move is probable that the prediction will place the target in the exact same position as before. Instead, if it moves the probability that the tracked object will be located close to the previously known position is higher than to be located far away. Specifically, the likelihood can be seen as a **Multivariate Normal Distribution** centred

at the previous position and stretched towards the direction where the subject is moving (Figure 4.1h).

- ➤ If multiple similar objects exist in the frame and the tracking is following one of them, it is probable to lose it. This happen with a fast movement that generates an unpredictable big shift, and the target moves far away from the previous position while the tracker recognise its subject into a similar object. From now on, the tracking is broken because it follows the wrong physical element.
- ➤ A big shift can also be wrongly interpreted as a total occlusion.
- ➤ On the other hand, reducing the tolerance to big movement allows the algorithms to only search locally around the last known position resulting in a big improvement of the overall performance.

- **Low-resolution images** (Figure 4.1a): as with many other computer vision challenges the resolution of the input image is fundamental. A low number of pixels per frame results into a bounding box (often a small portion of the entire image) of very low resolution. Because of this, recognising the key elements that identify the subject is hard, but also faster and computationally cheaper.

- **Moving background/camera**: working on a fixed camera allows the use of a set of techniques based on background subtraction. The key idea is to get prior knowledge of the background and understand which objects are there, by removing the known part pixel by pixel. A big change of luminosity can make this harder, but still possible.
  A camera that rotates always on the same section can be managed as a fixed camera. This can be done by combining all the images along with the rotation as a unified one, creating a panorama image.
  Instead, a moving camera implying a moving background makes everything harder:
  - No background subtraction can be done.
  - The subject may change aspect even if it does not move.
  - The occlusion, both partial and total, can occur more easily.
  - A zoom or rotation of the camera causes a similar effect such as the "fast-moving object".
  - To estimate the movement speed and direction of the subject it is necessary to know the movement of the camera because one is relative to the other. However, almost always the displacement of the camera in the space is unknown.

- **Real-time**: designing a method to run in real-time requires to focus on the computational capability and respectively to the processing speed. Real-time might vary from 1 to 60 FPS according to the application. To achieve this speed it is often necessary to choose a faster method instead of an accurate one. This reduces the overall precision of the entire designed algorithm.

- **Long-term video sequences**: the input frame sequence is longer than a few seconds, up to minutes. Firstly, a long video might easily contain some of the problems listed above. In addition, trying to locate the same subject over and over

again without a reinitialization can fail due to the drift problem.
The **drift problem** consists of an accumulation of small errors along the tracking period. A tracking algorithm refers to the bounding box generated for the previous frame. The subject that should be located again in the new frame is extracted from that bounding box. If the tracking lasts for a small number of frames, the subject may look similar to the original one. Instead, during a long processing the bounding box starts to derive from the original subject. Due to partial occlusion, the box might be reduced. Due to a strong change of luminosity, the bounding box might be misaligned. Due to a fast movement, the box can be linked to the background. By continuing summing up all these little problems the main subject will be recognised as a superfluous element. Therefore the tracker stops to consider it useful and the track fails.
To solve this problem the solution is to re-initialize the tracker often before reaching the limit of the drift. This can be done with different methods such as a **Kalman filter** or by recognising the main subject with a detection, that is the method presented in this thesis.

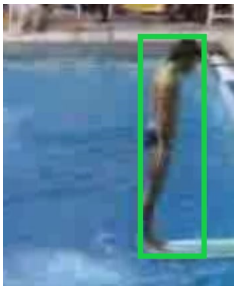
- **Multiple subjects** (Figure 4.1e): if there is more than a subject to follow the naive solution is to apply an object tracker to each of them. Each tracker works independently from the others therefore given X subjects the performances f measured in FPS are reduced to f/X.
An efficient solution consists of considering all the tracked subject at the same time with a single algorithm instance.

- **Type of the subject** (Figure 4.1c): the baseline problem does not assume any kind of prior knowledge about which type of subject should be tracked. However sometimes ad hoc solutions are required, which simplify the problem. Frequent choices are people (such in this thesis), animals, vehicles or inanimate objects. For example, if vehicles are chosen the change of shape is not a problem because a car always looks the same.

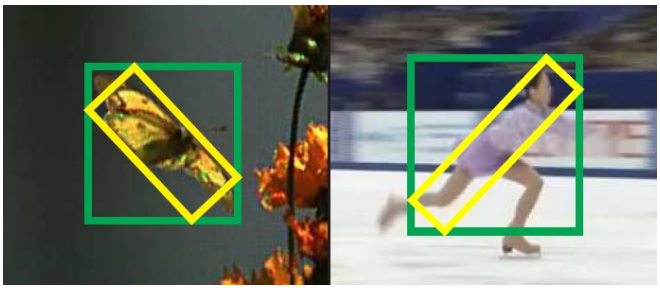
4.1.4 The traditional tracking problem compared to the thesis project

Some of the aspects explained in the previous section are extremely frequent in a lot of videos and databases, hence they are considered as a normal scenario. Others instead are often not even considered in the samples and the majority of the algorithms do not officially solve them.

- The conditions that are often respected in object tracking are:
  - The low-resolution images to allow the elaboration even without powerful computation devices, such as a smartphone.
  - A moving camera because a great majority of videos are in movement and shaking, except for the video surveillance field where the video camera is placed.
  - The changes of 2D shape and the fast-moving objects should be respected, the robustness on these two points make the algorithms more or less reliable.
  - A short-time total occlusion might sometimes be managed, but it is rarely guaranteed.



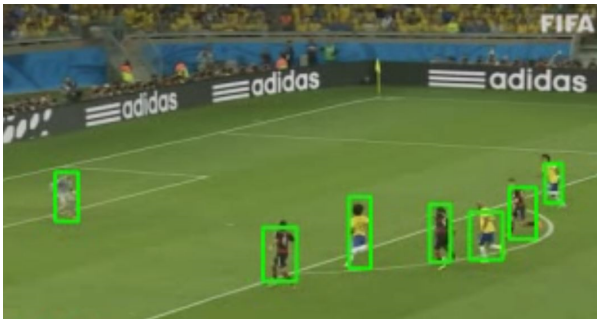
(a) Images with low resolution.



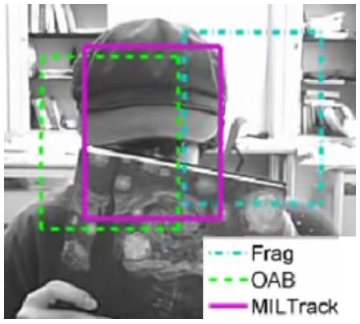
(b) Two subject that can vary their appearance very quickly.

(c) Tracking deals with objects, humans, and even animals.

(d) The limbs disappears.

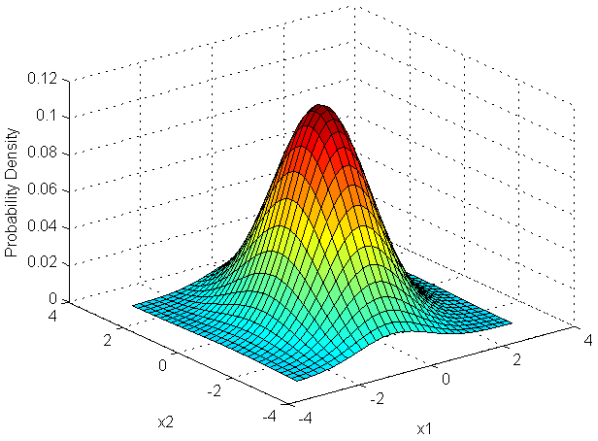


(e) A soccer match action where all the players are tracked.



(f) A face partially hidden by a hat and a book.

(g) The new location is close to the older one.



(h) The probability of where the new location is respect a multivariate normal distribution.

(i) A motorcycle running on a street, is totally hidden by a tree for multiple frames.

Figure 4.1: Some visual examples of the tracking challenge conditions.

- The requirements for this thesis, except the traditional ones, are:
  - Long-time occlusion. In our scenario where the robot physically follows a

person, the Leader can disappear behind a corner and will be hidden as long as the robot reaches it.

- Real-time processing. We have decided that to understand the movements in the real environment is sufficient a processing speed of 5 FPS.
- Long-term video. The algorithm is designed to last for a very long period, no explicit bounds exist since the drift problem has been solved.
- Subject limited to people. We are not interested in following animals or vehicles, even if extending the algorithm to them only require to change the object detectors setup and the internal database of images to train the recognition procedure.

4.2 Principal known algorithms

In this section a set of algorithms that perform well to solve the tracking task is presented. Differently, from object detection, the number of existing methods for tracking is much wider. This happens due to the high variability of the problem. The methods shown below represent a trade-off in terms of speed and reliability.

4.2.1 MIL (Multiple Instance Learning) tracker

The MIL tracker[21][22] is an extension of the older **BOOSTING Tracker**[23], both methods are based on an **online classifier**. "*Online*" means that the classifier is trained "*on the fly*" during the execution of the algorithms and not in advance. This type of training does not allow to use thousands of images but very few. An application of this method is presented in Figure 4.2.
The idea of the online classifier is to trust the initialization of the tracker and to use the initial bounding box as the first training sample. The negative samples are then generated taking rectangles that do not overlap the positive example. The classifier learns during the execution to recognise the tracked subject as positive and the rest as negative.
The frames after the first one are elaborated similarly. The positive subject is searched around the last known position and the classifier assigns a probability to each proposal. The box with the highest score is chosen as positive and it is used to continue the training of the classifier.
The novelty of MIL compared to BOOSTING is shown in Figure 4.3).
Instead of using only the positive sample to fit the online classifier, MIL creates a bunch of bounding boxes proposals around the positive sample, called **bags**. All these boxes should contain the subject and one could even be perfectly centred on it. The training is done with the "Multiple Instance Learning" that takes the bag of positive proposals and selects the best one (the more centred one) to improve the classifier. In the end, the instance is trained with only one box that was chosen starting from a set of good alternatives and not with all of them. The negative samples are then generated as for the first frame.

4.2.2 KCF (Kernelized Correlation Filters) tracker

The KCF tracker[24][25] is an additional extension of MIL tracker.
The key idea is that the sampled images are similar due to the subject that is repeated or thanks to the background that does not change extremely fast compared to the FPS of the video. This repetition of similar patterns can be used to optimize the operations and speed up the computation. The technical improvement comes from the application

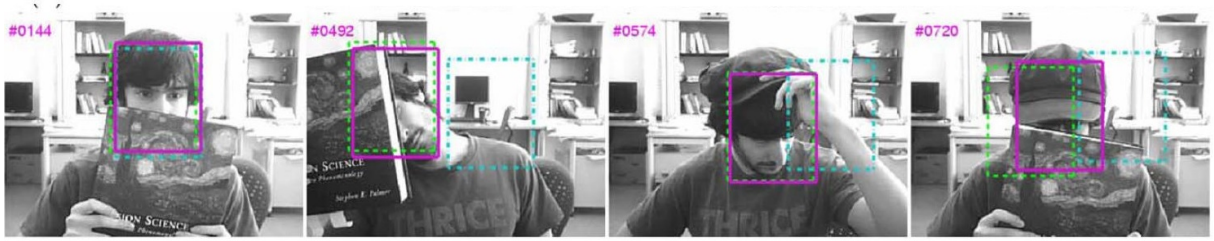


Figure 4.2: A sample of the MIL classifier tracking the face of the author while is partially occluded by a book. MIL tracker is in purple, while cyan and green are respectively FragTrack[1] and Online Ada-Boost[2].

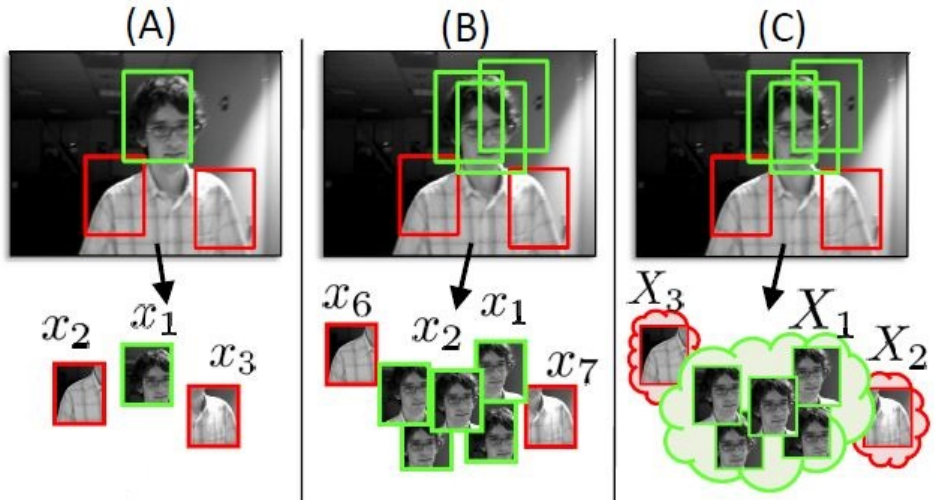


Figure 4.3: A comparison of how positive and negative samples are selected, to train the online classifier. In A the selection of BOOSTING tracker, the positive sample is used. In B a set of proposals generated around the positive are all used to train. In C the selection of MIL tracker, bags of samples are used and only one proposal is extracted and accepted from each one.

of the **FFT (Fast Fourier Transformations)**, which allows to apply the elaboration of images in an efficient manner.
The novelty introduced with KCF allows this algorithm to outperform both BOOSTING and MIL, in terms of accuracy and speed. The weakness of this chain of three methods is the full occlusion. None of them is able to deal with total occlusion that always causes the tracking failure.

4.2.3 Median Flow tracker

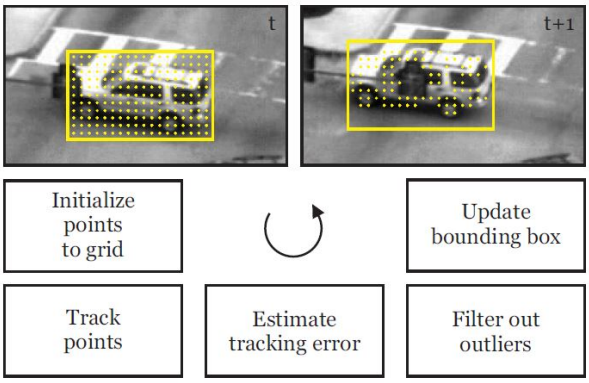
The Median Flow tracker[26] is a reliable method that locates the subject according to its trajectory. The key idea is that the algorithm recognises points in two subsequent frames. These points should be the same physical element in the real space.
The overall procedure is shown in Figure 4.4. The first step (Figure 4.4a) consists of the creation of a grid of points on the initial bounding box, and then the localization of these points in all the future frames. This connection through the frames helps to know the exact motion model of the tracked algorithm. The **Motion Model** (MM) is the combination of actual position (x, y) and velocity. It is defined with the angle or direction of motion (θ) and the module of the speed (s). Essentially knowing how the subject is

moving helps to predict where it will be in the near future.

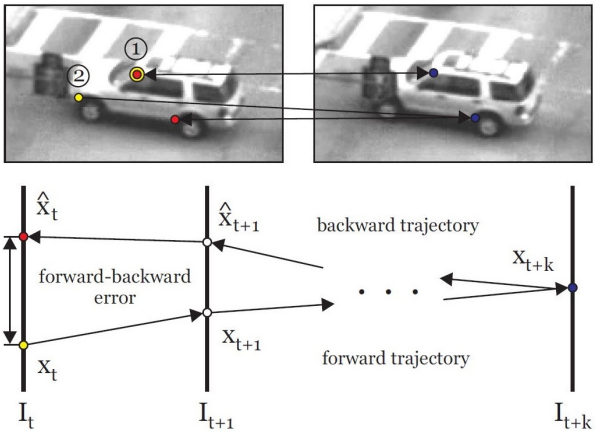
$$MM = ((x, y), (\theta, s))$$

The interaction of the frames works as follows. Every time that a new frame is added, the knowledge of the motion model suggests where the subject can be located. Then, the key points are searched in this new image. Once they are found it is fundamental to check the consistency of the trajectories (Figure 4.4b). Each new point is associated with the most similar already known physical point and vice versa. If this double matching is correct and the two points refer to the same physical element the trajectory is confirmed (point 1 in the figure), otherwise there is a misalignment in the trajectories (point 2 in the figure). In Figure 4.4b point 2 is firstly (forward pass) associated with the front wheel, in the frame after is linked to the back wheel since the other one is hidden from the street signal. The backward pass links the back wheel again on itself. This misalignment of the connection is recognised as an error and so the point cannot be trusted. If no point can be trusted the tracking fails.

Thanks to this double-checking procedure, the Median Flow tracker is a method that is able to well recognise when the tracking has failed to follow the subject. Unfortunately, the requirement to match key points over and over in the frames reduces the capability of the algorithm to manage scenarios where the subject appearance change too much. For this reason, this algorithm is not good for tracking high deformable subjects such as animals and humans.



(a) The first bounding box is divided into a grid of points, and a portion of them are recognised in the next frame. Below, the main steps of the algorithm.



(b) The points found in two consecutive steps are compared to check if the forward and backward trajectories are the same and then it can be trusted.
Point 1 is accepted while point 2 is rejected.

Figure 4.4: The overall procedure of the Median Flow tracker computing forward and backward trajectories to precisely locate the target.

4.2.4 CSRT (Channel and Spatial Reliability Tracker)

The full name of the method is Discriminative Correlation Filter with Channel and Spatial Reliability (DCF-CSR)[27]. Such as KCF (Section 4.2.2) and others, even this method is

based on correlation filters.

A **cross correlation filter** is a technique that aims at localizing into an image the exact position of another one. A representation of the cross-correlation usage done by CSRT is shown in Figure 4.5. In details, the portion of the last frame, delimited by the last known bounding box, is elaborated with multiples correlation filters, each one producing a different output. These elaborations simulate possible changes in the appearance of the subject. Designing good filters is fundamental to well match the variability of the tracked subject and the generated features. The filter outputs are modified images of the last bounding box cropped area.
The output of each filter (an image) is moved along the full picture pixel by pixel (learning stage: Figure 4.5 left) to check which portion of the entire camera view is more probable the subject that we are looking for. The result of this scan is a confidence map that should present a peak in correspondence of the new position of the tracked subject.
All the filters can then be summed up together (localization stage: Figure 4.5 right) to highlight the proposal of each one and comes out with the final response. This response shows exactly where the subject is placed in the new frame. A visualization of a confidence map applied to the original image is shown in Figure 4.6.

The characterization of CSRT is focused on the type of correlation filters used, with the idea of using a lot of them and combine the results at the end to produce a more reliable localization. The very high accuracy that this algorithm offers is compensated by the low FPS rate that it achieves.

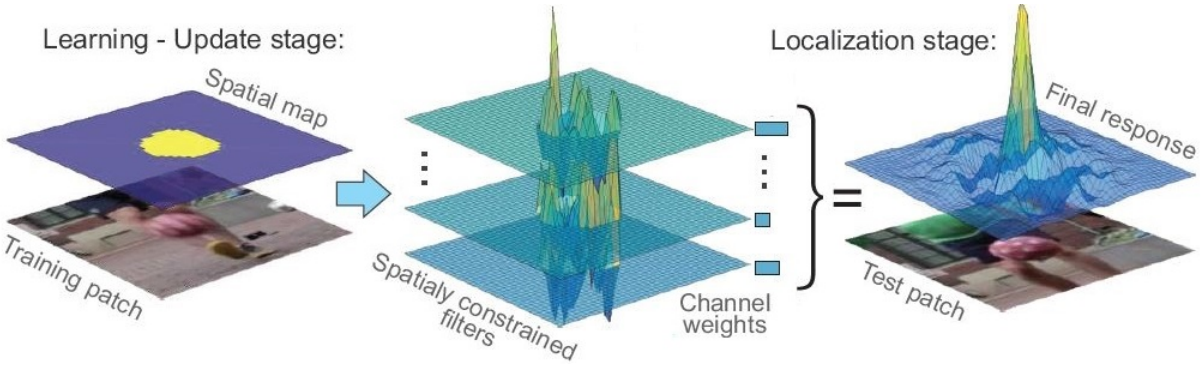


Figure 4.5: The overall procedure of the CSRT algorithm. The learning stage is composed of multiple correlation filters that are applied to the input frame. Each filter produces a confidence map that highlights where the subject should be. The localization stage combines all these confidence maps to produce the final response that precisely locates the subject.

4.2.5 MOSSE (Minimum Output Sum of Squared Error) tracker

The MOSSE algotihm[28] such as KCF (Section 4.2.2) is a method whose strength lies into mathematical smart choices instead of a complex high-level logic like MedianFlow (Section 4.2.3).
The novelty is introduced with a new correlation filter, called MOSSE. It can be applied to the input frames and precisely locate the variations fundamental to understand the movement of the subject.
Despite the original article states robustness against variations in lighting, scale and non-

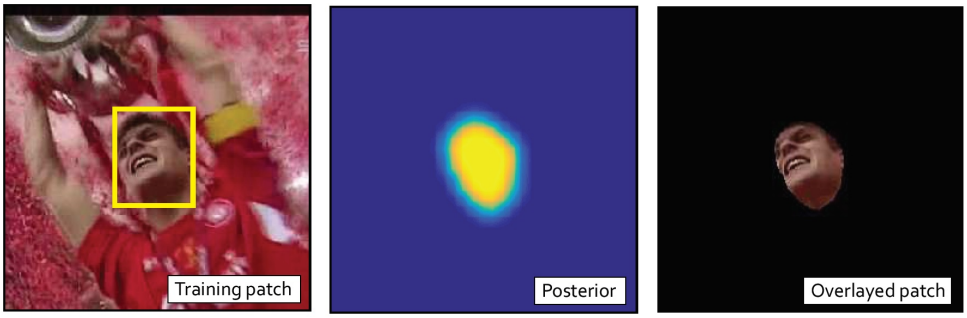


Figure 4.6: A visualization of the confidence map applied to track the face of the man in the image. On the left the original image, centre the confidence map in 2D. On the right the cropped image according to localization.

rigid deformations, this algorithm is not so reliable as it appears. On the other hand, the strength of this method is the extremely fast computations (FPS rate) that outperform all the other trackers presented in this section.

4.2.6 GOTURN (Generic Object Tracking Using Regression Networks)

GOTURN[29] is a tracker based on neural networks. Differently from MIL and KCF (Section 4.2.2), this method is not based on an online NN, such as the online classifier, but it is based on an offline CNN.
An **offline NN** is a traditional NN that is trained on thousands of data: couples of images, in this case, producing the trained model. The process is done in advance, before the effective use, and not "on the fly" while running the tracker. The generated model is used at run time to know how to respond to an input value. The big advantage of offline methods is that the training procedure is the slowest section, for this reason an online classifier could not perform at very high FPS rates.

The general workflow of the tracker is shown in Figure 4.7. The tracker takes as input a frame at a time and always compares it to the previous frame. This choice simplifies and standardizes the input to empower the potentiality of the CNN. However but makes the algorithm to have no chance against total occlusions even for one single frame. The CNN takes as input two squared images cropped from the frames. The crop on the previous frame is a bounding box centred around the last known position with some margins that will contain even the location in the frame afterwards. The current frame is cropped based on the same bounding box, but in this case, the subject is not centred in the square.
Then, both squared images are processed with two independent stacks of convolutional layers, followed by three fully-connected layers. The final output is composed of four values representing the top-left and bottom-right corners of the bounding box. It is centred on the subject in the squared image of the current frame. Some examples of the application of the CNN are shown in Figure 4.8.

4.2.7 TLD (Tracking-Learning-Detection)

Differently from all the other methods, TLD[30] aims to be a complete tracker able to deal with extremely complex scenarios. If the previously presented trackers have no chance to manage a long-time total occlusion and a long-term tracking, this method is able to overcome both of the problems. A sample showing potentiality is in the Figure 4.10.

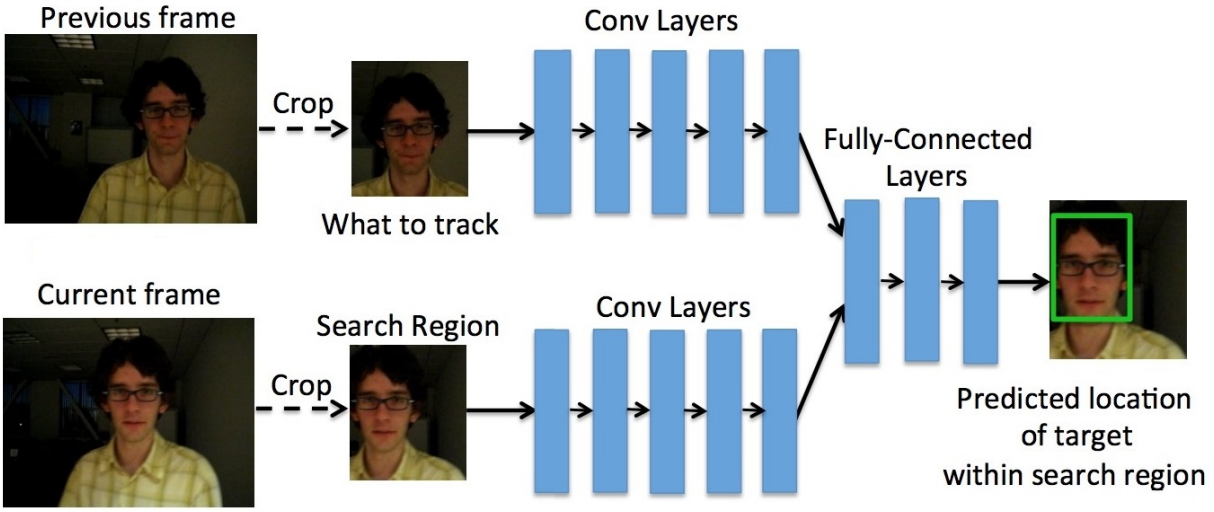


Figure 4.7: The overall procedure of the GOTURN algorithm. On the left, the frames coming from the camera are cropped with the same square, that is centred on the subject in the previous frame (above). The images are elaborated with CNN to produce the output bounding box, that highlights where the subject is in the current frame.

The foundation principle is that TLD is not a single algorithm but it is a combination of three. The interaction of these three parts is shown in Figure 4.9. The key idea to overcome the re-identification problem that occurs after a total occlusion, or the drift problem that happens along a large video, is to often re-initialize the tracker. In fact, the tracking (**T**) of TLD aims at managing short-term video clips. When a small problem occurs the detection (**D**) tries to locate the subject back again. While these two situations take turns, the learning procedure (**L**) extracts the key elements that recognise uniquely the subject and understand how to precisely locate it inside the frame.

The trade-off to use this extremely flexible structure is paid with a not high FPS rate. However, the biggest problem of this method is the huge quantity of false-positive predictions. The learning procedure starts with a few samples, meaning that errors in the beginning phase can occur easily. The failure in the first phase causes wrong learning that will produce more and more errors later on.
Despite the good potentialities, this algorithm is not reliable for the traditional tracking task.

4.3 Which tracker could be chosen

In the previous section we have presented seven different tracking algorithms, each one with his own strengths and weaknesses. These methods are the ones that were explored during this thesis work, but many others exist.
The tracking task that we aim at solving is a long real-time sequence with long-lasting total occlusions, as explained in Section 4.1.4. This task is solved with a combination of detection, tracking and recognition, so the internal tracking challenge is much simpler.
The requirement is to solve the baseline tracking task (Section 4.1), and to deal with changes of shape and partial occlusions. Considered the explained methods and known their speed performances (shown in Table 4.1), here it is what we have chosen. Note that

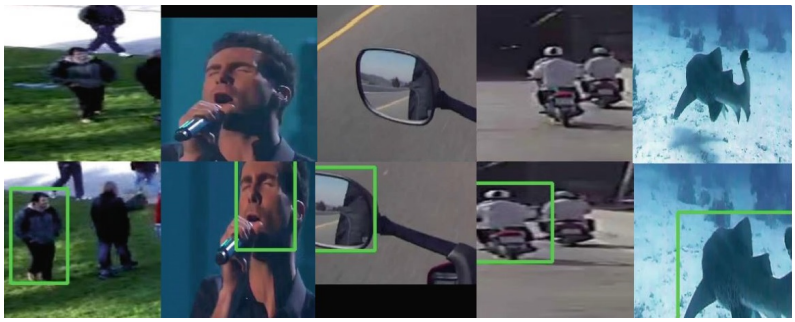


Figure 4.8: Samples of application of the CNN of GOTURN. The two squared images coming from previous (above) and current (below) frame are fed into the network. The answer is the green bounding box that locates the tracked subject on the new frame.

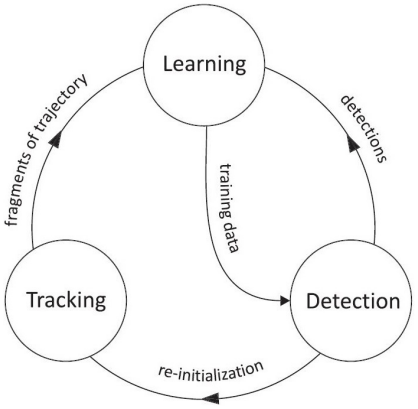


Figure 4.9: The interconnection of the three foundation methods of TLD algorithm.

Figure 4.10: A sample that shows the potentialities of the TLD algorithm, dealing with a total occlusion while tracking a motorbike with a dramatic change of size. The red dot means that the subject cannot be found.

the underlined methods are the best trade-off choices.

- **MIL tracker** is not a good choice because it runs at few FPS and a newer version that outperforms both its speed and accuracy exists.
- **KCF tracker** is the new version of MIL. It is one of the fastest methods and it is also reliable. It suffers the rapid change of appearance and, less important for our scenario, it does not manage total occlusions.
- **Median Flow tracker** works well only on not deformable or rigid subjects. Since we are dealing with humans it is a bad choice.
- **CSRT** is the most reliable method, it could even manage short total occlusions. Despite the low FPS rate is a great choice, in fact, our goal is a tracker running at around 5 FPS that is way less than the speed of CSRT.
- **MOSSE** is focused on pure speed. It is not the most reliable method but it can be a good choice to try to reach the highest FPS rate for the general challenge.
- **GOTURN** is a reliable method, running at a good FPS rate. At the moment it is not integrated in this project but it can be a great choice for future improvements.
- **TLD** is a too complex method. It is able to solve long total occlusions but easily fails on simpler scenarios, by proposing a lot of false positives. It is not reliable at

Algorithm	MIL	KCF	MedianFlow	CSRT	MOSSE	GOTURN	TLD
FPS	9	38	40	15	56	20	10

Table 4.1: An overview of the FPS rate of the trackers described. The performances were measured on an Intel Core i5 CPU and on an Nvidia Jetson TX2 GPU. The speeds measured are almost the same.

all for our purpose.
TLD is based on a principle that is similar to the one proposed in this thesis. The integration of tracking and detection linked together with a third procedure: learning in case of TLD and person recognition for this project. Both algorithms aim at solving the total occlusion problem and the drift problem with a reinitialization of the tracker performed with an object detector. The key difference is the existence in this thesis of the slow start phase presented in Section??? TODO.

5 People Recognition

The goal of this chapter is to shows which types of techniques could be used to understand, given two images of a human, if the represented person is the same or not.

5.1 Problem definition

The **people recognition module**, also known as the **people identifier module**, overall the entire project as the role of connecting in a consistent way the two remaining modules, the detector and the tracker. The tracker works for the majority of the time, following the **main subject**, called **Leader**. Then, when it fails or after a certain amout of time, the detector locates all the people in the incoming frame. Finally, the recognizer has to answer a simple question:

> *Is this person the Leader?*
> Formally, given an image, cropped on the bounding box of a person, and a set of other images, with the same characteristics, containing various people labelled with names. Choose which is the name of the unknown person.

In our specific scenario, we are interested only into understanding if a bounding box contains a representation of the Leader or not. For this reason the label can be considered with only two values: "*positive*" or "*negative*", "*Leader*" or "*other*".

5.1.1 Video surveillance application

This challenge is extremely important in the video surveillance field. In fact, one of the most common application is to reconstruct where a person was seen during a certain time slot. The video surveillance application is a little bit different from the one in which we are interested:

> Given a dataset of images of people, and given as request a query containing an unknown image of a person, extract from the dataset all the images that contain the same person of the query.

From a high-level point of view, the two problems might look different but essentially are the same. The key idea is to look for all the images that might contain the person of the query. Then, according to the task, retrieve this list of people or use them and all the associated information to understand who is the person of the query.
Another difference comes from the acquisition of images. In our scenario, the webcam is mounted on the robot that is moving around, while in video surveillance the cameras are often static but multiple of them can simultaneously collect frames. We took inspiration from both single-camera scenarios[31] and multi-camera systems[32].

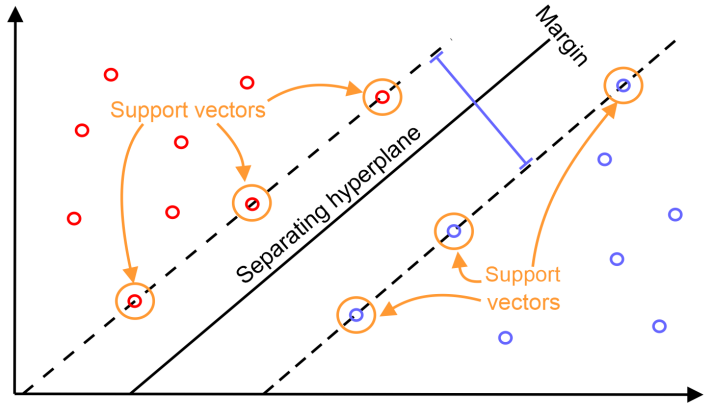


Figure 5.1: The overall mechanics of SVM. The red and blue classes are divided by the hyperplane that maximizes the margin length, defined as the shortest distance from the closest point on both sides.

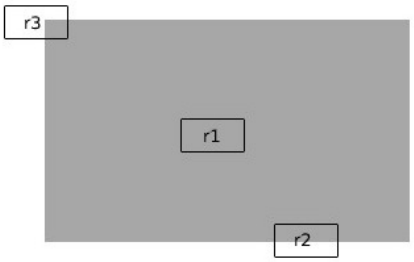


Figure 5.2: Three possible regions that can be used as keypoints. R1 cover a complete flat region. R2 is located on an edge. R3 is placed on a corner, this is the location that can be better recognised.

5.2 Not working methods

This section is focused on some unsuccessful technique was explored during the development of the general structure for the project.

5.2.1 Offline algorithms

An offline algorithm (explained in Section 4.2.6) is a great choice for training methods able to judge situations quickly. In this case, we have considered an **SVM (Support Vector Machine) classifier**.
This method is a supervised learning model based on a linear separator. The essential idea is to find the line[1] that divide the dataset into two sections where respectively the two classes of interest lie (a visual representation is given in Figure 5.1). The method could work if each image of a person is converted into a point, as explained in Section 5.3.1.
But this approach has a ground truth problem. The offline methods require prior knowledge of the context of application to train ad hoc models. In our case, the element to know is "*Who is the Leader?*". Based on this information the dataset can be split into two classes, the images representing the Leader and the other ones, then the SVM classifier can be trained. The result is a model that always recognise one person, exactly the same person all the times.
Unluckily, this prior knowledge is not given. The algorithm should recognise all types of people after seeing them for a few seconds. This is the context where an online algorithm can be wisely used.

5.2.2 Key points matching

The key points, or **feature points**, matching is a technique that compares two images and tries to recognise which are the common elements of the two. It is widely used to find small patterns in complex images, understand the variation of orientation, stitching panorama view and judge if the subject of two pictures could be the same.
This technique is based on key points. These elements are position in the image that can be easily located and identified as unique. In Figure 5.2 are shown three regions, a flat area, an edge and a corner. The most recognizable point is the one placed at the

[1] A line in 2D problems, that is converted into a hyperplane in N-dimensional scenarios.

corner. If an object appears in two images a key point should be recognised in both the appearances.

Technique definition and known algorithms

This technique is divided into two modules:

- **Feature localization**: This module working on a single image at a time, is internally divided into:
  - **Feature detection**: Given the raw image the goal is to identify all the corners that can be easily re-identified in another image.
  - **Feature description**: A small area containing few pixels cannot be easily matched. So, the descriptor needs to take each identified key point and normalize them in order to be independent respect to the image conditions like illumination and orientation. This key point enriched with additional information is the feature point.

  A visualization of the feature localization is shown in Figure 5.3.
  To solve this task exist several methods were used[2]:
  - SIFT (Scale-Invariant Feature Transform)[33] is able to manage both rotations and scale variation of the feature points.
  - SURF (Speeded-Up Robust Features)[7] is the advanced version of SIFT with all its pro but, in addition, has a lower processing time and produce an extremely high number of feature points.
  - ORB (Oriented FAST and Rotated BRIEF)[34] is the only algorithm of the three that is not patented. It is a combination of two sub methods:
    - ➤ FAST (Features from Accelerated Segment Test):[35] method completely focused on pure speed.
    - ➤ BRIEF (Binary Robust Independent Elementary Features)[36] algorithm extremely sensitive to noise, use Gaussian kernels to remove it and work precisely.
- **Feature matching**: The third module of this technique works on two images after that the feature points were computed. The goal is to match the features of the two images together. A match can only be one by one. A feature point should have exactly one corresponding point in the other image and vice versa, to be a correct and reliable connection. If a point is linked to more than one on the other image, this means that the connection works with multiple inaccurate key points hardly recognisable.
  In Figure 5.4 is presented an example working on two similar images. All the correct matches, draw as green lines, can be used to both assigns a probability that represents, how likely the subjects are the same "object". But also to understand which type of 3D transformation can be applied to an image to convert it into the other one.

  The existing algorithms for the matching are:

[2]Note that since the mechanics of the methods are similar and this technique has not actually been implemented in the thesis the algorithms reported here are not explained.

– BFM (Brute Force Matcher) This is the naive solution. All the possible combinations of points of the two different images are tried. Obviously is a precise method because it does not work under assumption but if the number of key points is high this method is at all infeasible.
– FLANN (Fast Library for Approximate Nearest Neighbors) matcher[37] Differently from BFM, this method is optimized to work even with a large number of features. The optimization is based on the local search of K-Nearest Neighbors.

Figure 5.3: The feature localization module applied to a Tesla. The key points are spread along the edges of the car and the background. Instead, the road has not, because the patter is often repeated, hence it is not reliable.

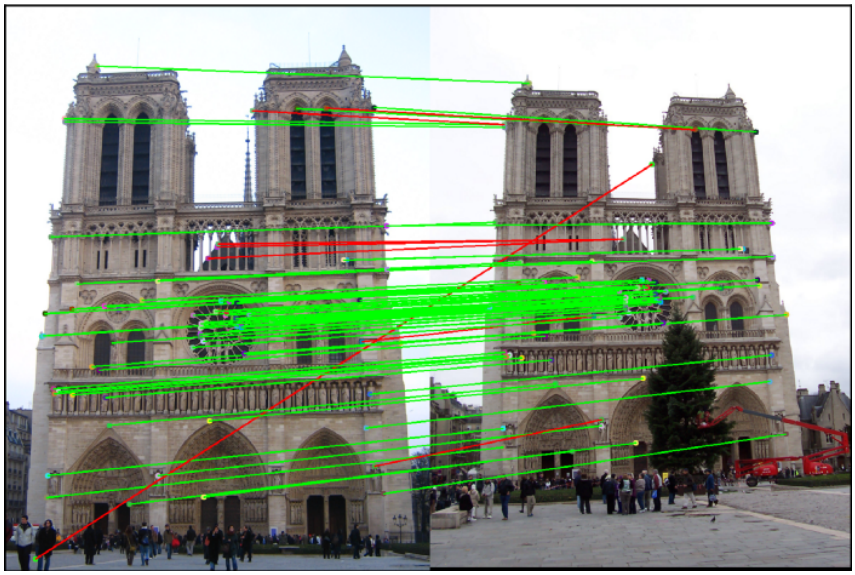


Figure 5.4: Two images of Notre Dame de Paris are aligned with the feature matcher. Since the image is very clear a lot of points are paired correctly (green lines), while a few of them are not (red lines).

Key points matching applied to people

The initial idea was to apply the key point matching to people. The test were done on the **PRID450 (Person Re-IDentification)**[39] dataset that contains thousands of images

of cropped people walking outdoors. The dataset is constructed with multiple shots of the same person on different moment and prospectives.
The idea has two big problems:

- The images cropped around the people has a very low resolution. The result is that the details that could distinguish a person from another one cannot be visible, or better cannot be recognised.
- Humans present a high deformable-body, with a surface (clothes) that continuously change aspect. Instead, the key point matching is designed for a pattern that is repeated often and clearly. The consequence of this is a matching that works as if it were random.

In Figure 5.5 are shown some examples that visually demonstrate the unreliability of this technique applied to humans.

Figure 5.5: Some matching samples show that key point matching easily fails under these conditions. There are multiple wrong aspects: people that present no key points, objects such as bags that have plenty of features, matching that connects completely different parts of the body like shoulders with legs. Even with the same subject with almost the same position (bottom-left images), the algorithm fails with most of the points. The exception is the bottom-right image that has a perfect matching, but the two pictures are exactly the same one. So it is not a reliable test.

5.3 KNN (K-Nearest Neighbors) with images into N-dimensional space

Since that offline algorithms cannot be used, the official solution is based on online algorithms. The chosen method is KNN (K-Nearest Neighbors) classifier is a widely used method that is based on the proximity of points into space, where each point has a class. When a new point should be judged, the K nearest known points are found, and the class for the new incoming is chosen according to the majority of classes of the K selected

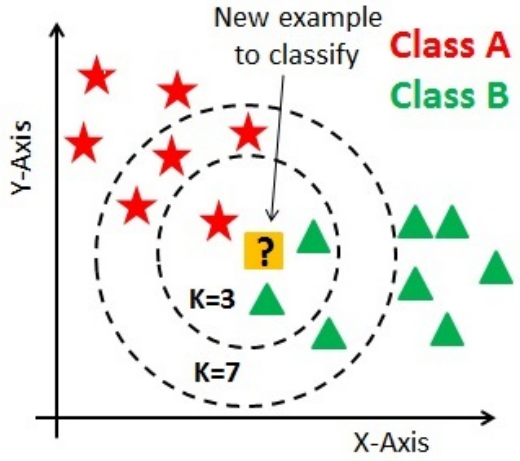


Figure 5.6: Example of the mechanics of KNN. The new point (question mark) will be classified as class B if the 3 nearest neighbours are considered. It will be classified as class A if K=7 is used.

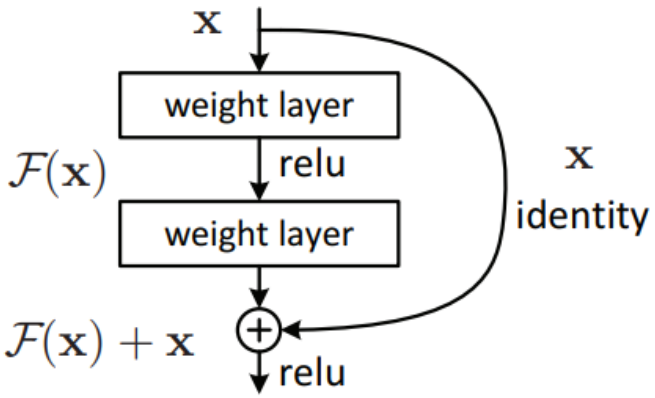


Figure 5.7: The residual block of ResNet. The identity connection allow to the output of the precedent level, to skip two convolution layers. The result $f(x) + x$ is the combination of the convolutions $f(x)$ combined with the identity x.

points (an example is shown in Figure 5.6).
About this algorithm, some aspect should be considered:

- KNN is a native online method. The difference between online and offline is about computing the training in advance or not, but KNN has no training phase. The training only requires to store the known points with the associated labels, and this is done at almost no cost.
- The elaboration works with points and not images, this remarkably reduces the computational classification cost.
  On the other hand, an image should be collapsed to a point, meaning that a good representation should be used to do not lose important information.
- The classification steps should compare the new cropped frame, transformed into a point, with all the other known points. So, a very high number of stored points will slow down the execution. But in our real-time scenario, this will occur only if the tracking will last for an extremely long period, and this will not happen.

To apply KNN, the challenging task of converting images into representative points should be solved.

5.3.1 Image classifiers for representative points from images

We have chosen to use CNN to create the representative points that should describe an entire image. Unlikely, it does not exist an explicit field of study that aims to create this kind of points. The solution relies into the adaptation of a widely explored machine learning challenge called **image classification** (Figure 3.2). The goal is to predict which kind of elements exist inside the picture.
The image classifiers based on CNN work as follows:

1. The input picture is resized to standard dimensions. In addition, colours and lights are normalized.

2. The image is elaborated with multiple blocks of convolution's layers.

3. All the features extracted with the convolutions are collapsed, with a "*flatten operation*", into an array of thousands of elements.

4. In the end, this vector is eventually reduced and the final predictions, one for each output class, are generated.

The goal is to generate a point from the input image of CNN. The fourth part collapses all the elaborated information into predictions that vary according to the context of the application. Instead, the third level produces an array: a list of N numbers that can be seen as coordinates of a point into an **N-dimensional space**. KNN works independently from the space dimensionality, so, it does not matter how many features are produced the CNN based algorithms.

The classifier chosen are the DNNs (Deep Neural Networks) **GoogLeNet** and **ResNet**[40]. A DNN has the capability to produce better results than a NN. On the other hand, the huge number of parameters used should be tuned during the training phase. This calibration of the values is extremely hard on small datasets due to the **vanishing gradient problem**. Therefore, both classifiers introduce a novelty aiming to solve the problem connected to the depth of the network.

ResNet (Residual Network)

ResNet[43] is build on top of the idea of "skip blocks of layers". The residual block is shown in Figure 5.7.
A "*plain CNN*" has convolutions stacked one after the other, in this case, there is an additional element: the **identity connection**. It means that no filters are applied, the input is shifted two layers down. This connection is used to propagate the information deep into the CNN without modifying them. The advantage is that the input image is preserved through the network and it is not affected by an elaboration that lasts for several layers (more than 100). This novelty is very important for small training sets that are not able to fine-tune all millions of parameters of the DNN.
The original paper comes out with multiple models characterized by different depths, for this thesis we have chosen ResNet50[40]. This model produces a representation point in 2048 dimensions.

GoogLeNet (Google Le-Network)

GoogLeNet[41] is based on a new convolution scheme called **Inception module**. The name and the goal of module comes from the quote "*We need to go deeper*"[3].
The naive version of the inception module parallelizes three different convolution filters (1x1, 3x3, 5x5) and a *max-pooling filter*. This special elaboration reduces the depth of the CNN while preserving its potentiality. On the other hand, the number of parameters is still huge. The official version of the inception module (Figure 5.8) stacks each filter (3x3, 5x5 and max-pooling) with a 1x1 convolution to reduce, by a factor of 10, the overall number of parameters.
We have used the model[42] proposed together with the paper. This GoogLeNet implementation produce a representation point in 1024 dimensions (half of ResNet50).

[3]Quote of the film Inception (meme)

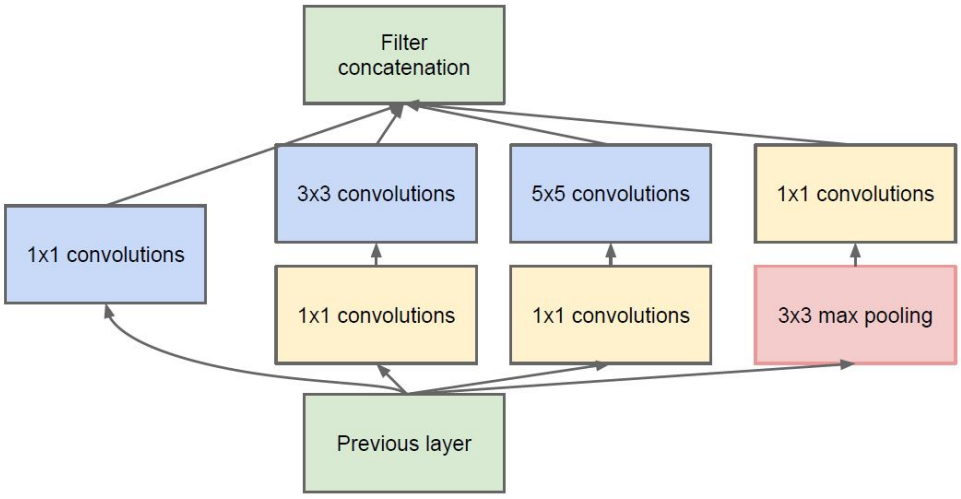


Figure 5.8: The inception module presented with the GoogLeNet model.

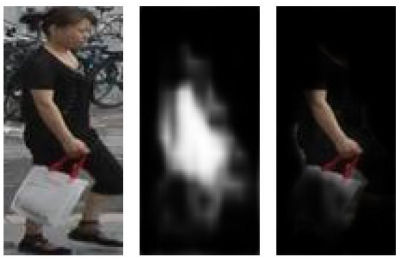
SSP-ReID (Saliency-Semantic Parsing Re-IDentification)

SSP-ReID[44] is a useful technique specifically designed for improving the potentialities of the image classifiers when are applied to bounding boxes of people. This method is ideated to create a representative array of features.
The model is based on a "*CNN backbone*" that can be selected from a wide set of pre-existing CNNs such as ResNet or GoogLeNet. The improvement comes from additional pieces of information that are processed together with the backbone.
The extra information generated (shown in Figure 5.9) are:

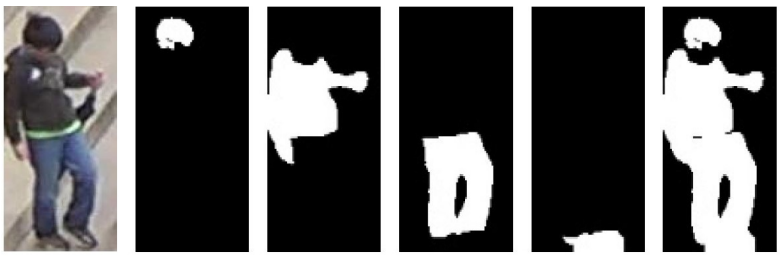
- **Saliency** (Figure 5.9a): this technology aims at isolating all the pixels that appear at "first glance" to the human eyes. The set of these pixels are the one that may highlight some key aspects of a person such as a bag, the clothes or other.

- **Semantic Parsing** (Figure 5.9b): the body of the person is divided into 5 sections and these are elaborated separately. Each body part can have its own characteristics. The 5 sections are: head, upper body, lower body, shoes and complete body.

This method is not integrated into the thesis project at the moment but it can be a great choice for future improvements.



(a) The saliency elaboration, applied to the woman, highlight the presence of a bag.



(b) The semantic parsing elaboration divides the body of the child in the 5 sections: head, upper body, lower body, shoes and complete body.

Figure 5.9: Examples of saliency and semantic parsing elaborations.

5.3.2 Examples of KNN applied to people recognition

The intuition of using KNN has been tested on the **Market1501**[38] dataset. Similarly to the PRID450 dataset, also this dataset contains sets of images captured from different prospectives and in different moments of hundred of people. Each real person has is own ID so different images of the same subject are associated to the same ID.
A couple of the results of the experiments are shown in Figure 5.10. This elaboration is done by selecting small datasets of 18 and 99 images, of 2 and 11 people respectively each one with 9 pictures per person. KNN was "trained"[4] with the representative points extracted from the images of the datasets. Then, the queries (images of people) were used to retrieve the most similar people. This was done by generating the representative points of the queries and for each one, the K closest points are retrieved together with the associated images.

It is important to focus on the ratio between correct and wrong responses, green and red respectively. In the test elaborated with ResNet50 (Figure 5.10a) there are almost 50% and 50% of wrong and correct responses, while in the test elaborated with GoogLeNet (Figure 5.10b) there is only one false prediction over 14. Despite the different algorithms used the results is independent of them.
In fact, the key difference is that in Figure 5.10a there are 11 classes, so $9 * 1 = 9$ samples of the correct person and $9 * 10 = 90$ of the wrong one. While in Figure 5.10b there are only 2 classes so 9 samples against 9. This different ratio between positive and negative training examples affect the result of the predictions.
In this thesis, we are dealing only with 2 classes: the Leader and the other people. Hence, we are interested in the scenario with 18 images that works extremely well.

Lastly, for the integration of person recognition, with the tracking and detection modules, we only need to know if the query belongs to a class or to another one. This choice is based on the most likely class on the first K[5] nearest neighbours of the query, so if the majority is green or red.

[4]The training of KNN consists of storing data and nothing more.
[5]In case of 2 classes K is often chosen odd.

(a) KNN applied on images elaborated with ResNet50. The training was done with 11 real people and 9 images of each one, for a total of 99 pictures. Here are shown 3 queries with the 8 most similar people.

(b) KNN applied on images elaborated with GoogLeNet. The training was done with 2 real people and 9 images of each one, for a total of 18 pictures. Here are shown 2 queries with the 7 most similar people.

Figure 5.10: In this picture are shown queries computed on the KNN classifier that has pre-processed small datasets of images of people. The query (top-left bounding box with blue contour) is used to extract from the database the most similar 7/8 pre-analysed people. The green contour means that the extracted person is correct, while if it is wrong the red is used.

6 Solution

7 Conclusions

Bibliography

[1] Amit Adam, Ehud Rivlin, and Ilan Shimshoni. Robust fragments-based tracking using the integral histogram. In *2006 IEEE Computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 798–805. IEEE, 2006.

[2] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-time tracking via on-line boosting. In *Bmvc*, volume 1, page 6, 2006.

[3] Dolomiti robotics. `https://dolomitirobotics.it/`.

[4] Sicong Jiang, Jianing Zhang, Yunzhou Zhang, Feng Qiu, Dongdong Wang, and Xiaobo Liu. Long-term tracking algorithm with the combination of multi-feature fusion and yolo. In *Chinese Conference on Image and Graphics Technologies*, pages 390–402. Springer, 2018.

[5] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[8] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[9] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855, 2006.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[17] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[18] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4620–4628, 2019.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[20] Wordnet graph. `https://wordnet.princeton.edu/`.

[21] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Visual tracking with online multiple instance learning. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pages 983–990. IEEE, 2009.

[22] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Robust object tracking with online multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1619–1632, 2010.

[23] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-time tracking via on-line boosting. In *Bmvc*, volume 1, page 6, 2006.

[24] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *European conference on computer vision*, pages 702–715. Springer, 2012.

[25] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van de Weijer. Adaptive color attributes for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2014.

[26] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Forward-backward error: Automatic detection of tracking failures. In *2010 20th International Conference on Pattern Recognition*, pages 2756–2759. IEEE, 2010.

[27] Alan Lukezic, Tomas Vojir, Luka ˇCehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2017.

[28] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2544–2550. IEEE, 2010.

[29] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 fps with deep regression networks. In *European Conference on Computer Vision*, pages 749–765. Springer, 2016.

[30] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1409–1422, 2011.

[31] Wiebe Van Ranst, Floris De Smedt, Jonathan Berte, and Toon Goedemé. Fast simultaneous people detection and re-identification in a single shot network. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[32] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018.

[33] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[34] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[35] Miroslav Trajković and Mark Hedley. Fast corner detection. *Image and vision computing*, 16(2):75–87, 1998.

[36] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer, 2010.

[37] Marius Muja and David Lowe. Flann-fast library for approximate nearest neighbors user manual. *Computer Science Department, University of British Columbia, Vancouver, BC, Canada*, 2009.

[38] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[39] Peter M. Roth, Martin Hirzer, Martin Koestinger, Csaba Beleznai, and Horst Bischof. Mahalanobis distance learning for person re-identification. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen C. Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 247–267. Springer, London, United Kingdom, 2014.

[40] Resnet50. `https://github.com/onnx/models/tree/master/vision/classification/resnet`.

[41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[42] Googlenet. `https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet`.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[44] Rodolfo Quispe and Helio Pedrini. Improved person re-identification based on saliency and semantic parsing with deep neural network models. *Image and Vision Computing*, 92:103809, 2019.