PhD Thesis

For the award of the degree of :

Doctor of philosophy of Polytechnic University Hauts-De-France

Discipline :

Electrical Engineering

Presented and defended by :

Amira MIMOUNA.

On 25*th* May 2021, in Valenciennes

Exploring Data Fusion for Multi-Object Detection for Intelligent Transportation Systems using Deep Learning

---

Committee

Reviewer:	Mrs. Dorra SELLAMI	Prof, University of Sfax
Reviewer:	Mr Youssef CHAHIR	MC HDR, University of Caen
Examiner:	Mr Abdeldjalil OUAHABI	Prof, University of Tours
Examiner:	Mr Jacques BOONAERT	MC, Institut Mines-Télécom Douai
Director:	Mrs. Najoua ESSOUKRI BEN AMARA	Prof, University of Sousse
Director:	Mr. Abdelmalik TALEB-AHMED	Prof, UPHF
Co-Supervisor:	Mr. Ihsen ALOUANI	MC, UPHF
Co-Supervisor:	Mr. Anouar BEN KHALIFA	MC, University of Sousse

---

Doctoral School :

ED-SPI

Research Laboratories :

IEMN-DOAE: Institut d'électronique, de microélectronique et de nanotechnologie, Département opto-acousto-électronique

LATIS- Laboratory of Advanced Technology and Intelligent Systems

Academic Year: 2020-2021

ABSTRACT

Building reliable environment perception systems is a crucial task for autonomous driving, especially in dense traffic areas. Researching in this field is evolving increasingly. However, we are at the beginning of a research pathway towards a future generation of intelligent transportation systems. In fact, challenging conditions in real-world driving circumstances, infrastructure monitoring, and accurate real-time system response, are the predominant concerns when developing such systems.

Recent improvements and breakthroughs in scene understanding for intelligent transportation systems have been mainly based on deep learning and the fusion of different modalities.

In this context, firstly, we introduce **OLIMP**[1] : A heter**O**geneous Mu**L**timodal Dataset for Advanced Env**I**ron**M**ent **P**erception . This is the first public, multimodal and synchronized dataset that includes Ultra Wide-Band (UWB) radar data, acoustic data, narrowband radar data and images. OLIMP comprises 407 scenes and 47,354 synchronized frames, including four categories: pedestrians, cyclists, cars and trams. The dataset presents various challenges related to dense urban traffic such as cluttered environments and different weather conditions. To demonstrate the usefulness of the introduced dataset, we propose, afterwards, a fusion framework that combines the four modalities for multi object detection. The obtained results are promising and spur for future research.

In short range settings, UWB radars represent a promising technology for building reliable obstacle detection systems as they are robust to environmental conditions. However, UWB radars suffer from a segmentation challenge: localizing relevant Regions Of Interests (ROIs) within its signals. Therefore, we put froward a segmentation approach to detect ROIs in an environment perception-dedicated UWB radar as a third contribution. Specifically, we implement a differential entropy analysis to detect ROIs. The obtained results show higher performance in terms of obstacle detection compared to state-of-the-art techniques, as well as stable robustness even with low amplitude signals.

Subsequently, we propose a novel framework that exploits Recurrent Neural Networks (RNNs) with UWB signals for multiple road obstacle detection as a deep learning-based approach. Features are extracted from the time-frequency domain using the discrete wavelet transform and are forwarded to the Long short-term memory (LSTM) network.

[1] https://sites.google.com/view/ihsen-alouani/datasets

The obtained results show that the LSTM-based system outperforms the other implemented related techniques in terms of obstacle detection.

Keywords: Intelligent transportation systems; Public dataset; Multi-modality; Fusion; Object detection; UWB radar; Entropy; Segmentation; Deep learning; LSTM.

RESUMÉ

Une perception fiable de l'environnement est une tâche cruciale pour la conduite autonome, en particulier dans les zones de trafic dense. La recherche dans ce domaine évolue de plus en plus. Cependant, nous sommes au début d'une voie de recherche vers une future génération de systèmes de transport intelligents. En effet, les principales préoccupations lors du développement de tels systèmes sont les conditions de la conduite, la surveillance des infrastructures et la réponse précise du système en temps réel.
Les récentes améliorations et percées dans la compréhension de l'environnement pour les systèmes de transport intelligents reposent principalement sur l'apprentissage profond et la fusion de différentes modalités.
Dans ce contexte, tout d'abord, nous introduisons **OLIMP:** A heter**O**geneous Mu**L**timodal Dataset for Advanced Env**I**ron**M**ent **P**erception [1]. C'est la première base de données public, multimodale et synchronisée qui comprend des données radar ultra large bande (ULB), des données acoustiques, des données radar à bande étroite et des images. OLIMP comprend 407 scènes et 47 354 données synchronisées, dont quatre catégories: piétons, cyclistes, voitures et tramways. L'ensemble de données présente divers défis liés au trafic urbain dense, tels que des environnements encombrés et des conditions météorologiques différentes. Pour démontrer l'utilité de la base introduite, nous proposons, par la suite, un framework de fusion qui combine les quatre modalités pour la détection multi-objets. Les résultats obtenus sont prometteurs et incitent à de futures recherches.
Dans les applications à courte portée, les radars ULB représentent une technologie prometteuse pour la construction de systèmes de détection d'obstacles fiables car ils sont robustes aux conditions environnementales. Cependant, ces radars souffrent d'un défi de segmentation: localiser les régions d'intérêt (ROIs) pertinentes dans ses signaux. Par conséquent, nous mettons en avant une approche de segmentation pour détecter les ROIs dans un environnement dédié à la perception de l'environnement c'est la troisième contribution. Plus précisément, nous mettons en œuvre une analyse d'entropie différentielle pour détecter les ROIs. Les résultats obtenus montrent des performances supérieures en termes de détection d'obstacles par rapport aux techniques de l'état de l'art, et une robustesse même avec des signaux de faible amplitude.
Par la suite, nous proposons un nouveau framework basée sur l'apprentissage profond qui exploite le réseau de neurones récurrents avec les signaux ULB pour la détection multiple

[1] https://sites.google.com/view/ihsen-alouani/datasets

d'obstacles routiers. Les caractéristiques sont extraites du domaine temps-fréquence à l'aide de la transformée en ondelettes discrète et sont transmises au réseau récurrent à mémoire courte et long terme. Les résultats obtenus montrent que le système basé sur l'LSTM surpasse les autres techniques implémentées en termes de détection d'obstacles.

Mots-clés: Systèmes de transport intelligents; Base de données public; Multi-modalité; La fusion; Détection d'objets; Radar ULB; Entropie; Segmentation; L'apprentissage profond; LSTM.

PUBLICATIONS

- A. Mimouna, I. Alouani, A. Ben Khalifa, Y. El Hillali, A. Taleb-Ahmed, A. Menhaj, A. Ouahabi, N. Essoukri Ben Amara, «OLIMP: A heterogeneous multimodal dataset for advanced environment perception», Electronics (Switzerland), Vol=9 , Date of Publication: 27 March 2020, https://doi.org/10.3390/electronics9040560, IF=2.412, SJR=Q2.

- A. Mimouna; A. Ben Khalifa; I. Alouani; N. Essoukri Ben Amara; A. Rivenq; A. Taleb-Ahmed " Entropy-based Ultra-Wide Band radar signals Segmentation for Obstacle Detection". IEEE Sensors Journal. https://ieeexplore.ieee.org/document/9316756 IF= 3.076, SJR=Q1

- A. Mimouna; A. Ben Khalifa; I. Alouani; A. Taleb-Ahmed; A. Rivenq; N. Essoukri Ben Amara " LSTM-based system for multiple obstacle detection using ultra-wide band radar". International Conference on Agents and Artificial Intelligence(ICAART 2021), February 2021, Vienna Austria, Class= C.

ACKNOWLEDGMENTS

I would like to acknowledge everyone who played a role in accomplishing this thesis.

To my beloved family that always supported me with their love and understanding. Without them, I could never have reached this current level of success.

I would like to give special thanks to my friends who have always encouraged me.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACF	Aggregated Channel Features
ADAS	Advanced Driver Assistance Systems
AI	Artificial Intelligence
AV	Autonomous Vehicle
AP	Average Precision
AWGN	Additive White Gaussian Noise
BPTT	Back-Propagation Through Time
Ca	Coefficients of Approximation
Cd	Coefficients of detail
CMDR	Constant Miss-Detection Rate
CNN	Convolutional Neural Network
DPM	Deformable Parts Model
DWT	Discrete Wavelet Transform
CA-CFAR	Cell Averging CFAR
CFAR	Constant False Alarm Rate
FCC	Federal Communication Commission
FN	False Negative
FP	False Positive
IOU	Intersection Over Union

IR-UWB Impulse Radio UWB

HOG Histogram of Oriented Gradients

HOS Higher Order Statistics

ITS Intelligent Transportation System

lidar Light Detection and Ranging

LBP Local Binary Patterns

LSTM Long Short-Term Memory

mAP mean Average Precision

MFCC Mel-Frequency Cepstral Coefficients

MOR Mapped Overlap Ratio

MFCC Mel-Frequency Cepstral Coefficients

MODP Multiple Object Detection Precision

MS-CNN Multi-Scale CNN

OLIMP A heterOgeneous muLtimodal dataset for advanced envIronMent Perception

PE Permutation Entropy

UWB Ultra Wide-Band

P Precision

R Recall

ResNet Residential Network

RF Random Forest

ROI Regions Of Interest

ROIs Regions Of Interests

RNNs Recurrent Neural Networks

RRC	Recurrent Rolling Convolution
R-CNN	Region-based ConvNet
SAE	Society of Automotive Engineers
SS	Selective Search
SSD	Single Shot Detection
SVM	Support Vector Machine
TP	True Positive
YOLO	You Only Look Once
1D	1 dimensional
2D	2 dimensional

1

INTRODUCTION

Contents

The main human capacities are the power and adaptability of our brain to process and interpret data acquired from our senses. While the human visual cortex is precise in processing the received information, humans can classify and locate complex objects that surround them by characterizing the objects' shapes, colors and orientations. Furthermore, humans are able to continuously analyze their environment and scan for new objects in their surroundings.
With the speedy ascension of Artificial Intelligence (AI), researchers aim to develop intelligent systems that efficiently imitate the human perception. To be accurate and reliable, these systems must consider several constraints that are spontaneously handled by the human brain. Autonomous vehicles (AVs) are one of the most important developed Intelligent Transportation Systems (ITSs) for the recent years. These vehicles require to perceive their environments like humans. Thus, the environment perception task is considered as one of the prominent fields of research.

1.1 THESIS CONTEXT

Traffic congestion has been increasing worldwide, which is caused mainly by the expansion of number of road users. This leads to a rising number of road accidents and fatalities, which definitely threats the quality of urban life.

Currently road injuries is estimated to be **the seventh leading cause of death** for all age groups globally, and it is ranked **first** for people aged between 10 and 24 years old [168]. According to the World Health Organization (WHO) [139], road traffic crashes generate a loss of over than **1.35 million** lives per year, where 54% are vulnerable road users. Moreover, they cause non-fatal injuries for almost 50 million people all over the world each year. From another point of view, these road traffic causalities account for 23% of all injury deaths worldwide in accordance to the distribution related to the global injury mortality via cause, as illustrated in figure 1.1 [121].
Furthermore, economically, road injuries engender considerable economic losses since they cost most countries around 3% of their gross domestic product [139].

Based on the statistics reported by the National Highway Traffic Safety Administration (NTHSA) [138] in 2018, **94%** of the road accidents are principally linked to *human errors*. These errors are caused mainly by fatigue, heedlessness and immature behavior (as immature driving, distraction when using the phone, etc.) according to information about crashes reported by the police [138].
For these reasons, AVs have been introduced in order to import an eloquent change for drivers, road users, infrastructure and pollution to contribute to fuel savings.

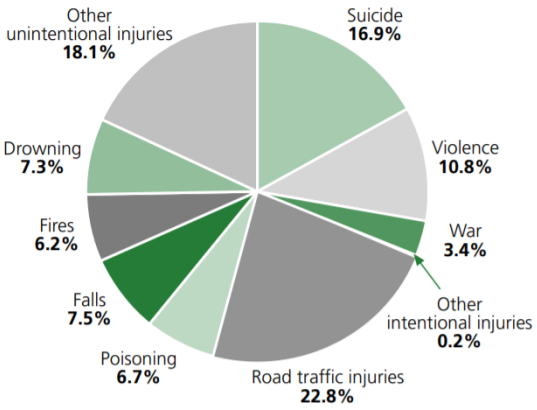


Figure 1.1: Distribution of global injury mortality by cause according to WHO Global Burden of Disease project [121].

Despite the fact that AVs aim to avoid accidents and provide economic, comfortable and intelligent driving, the development of such systems is highly complex. This is due to the permanent changes of the vehicle's environment that includes weather and illumination conditions, as well as the type of the obstacles and the behavior of road users.
However, with the recent improvements and breakthroughs of AI, which relies on the appearance of machine and deep learning algorithms, as well as the availability of new sensor modalities, AVs achieve huge progress. Thus, sophisticated driver assistance technologies have been developed to ensure road safety [148] such as collision avoidance, driver assistance and driver behavior monitoring systems[74] [75] [81].

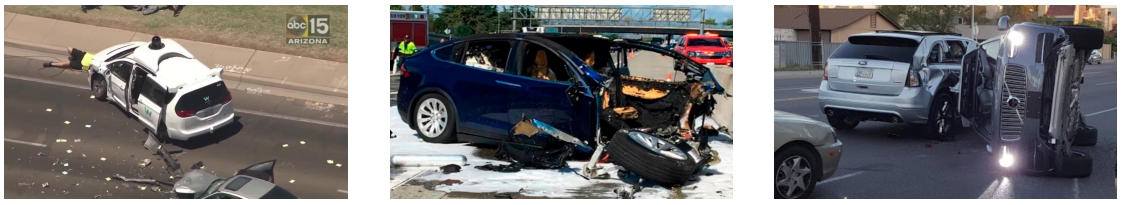
In fact, most of the proposed solutions require robust environment perception, especially in urban circumstances, and rely on the object detection process to prevent accidents and to protect, inter alia, vulnerable road users such as pedestrians and cyclists.
To achieve such requirements, object detection exploring sensors' fusion remains essential because the AV perception capability can exceed that of human drivers particularly in degraded conditions.

As the employed sensors offer complementary data and their collaboration can guarantee a better scene understanding and since training deep learning algorithms requires huge datasets, developing accurate and robust object detectors is a primary challenge in the AV field.

1.2 MOTIVATIONS AND OBJECTIVES

A long list of companies are interested in developing AVs and aim to launch highly automated vehicles by improving software capabilities and safety, like Audi, Google, Bosch, Nvidia, etc.

Despite the fact that AVs have shown immense progress during the last few years, there have been 13 serious car crashes using the autonomous mode or the autopilot mode [35]. Some of these crashes are shown in figure 1.2. On February $14th$, 2016, the Google's AV commits its first crash with a bus while lane changing [32]. One of the most critical of these serious accidents took place when the Tesla's Model X car driving in an autopilot mode had a collision with a highway barrier in California in 2018 [34]. This collision resulted in killing the driver. Accordingly, with the implication of Tesla with a number of other fatal crashes, the National Transportation Safety Board (NTSB) declared that both autopilot and driver errors were factors in these car accidents. Also, the Hyundai self-driving car crashed because of rain [33].



Google's car Tesla's car Uber's car

Figure 1.2: Serious crashes caused by AV.

Based on these reported fatal crashes that involve self-driving cars, the autonomous driving in urban roads remains an open and a challenging problem. In fact, the environmental variables, which vary from weather conditions to the surrounding human behaviors, are extremely indeterministic and difficult to predict.
For these reasons and since system failures lead to catastrophic accidents and fatalities, the improvements of object detectors and the development of new algorithms are still an inevitable process.

For the detection stage, most of the companies use either a unique sensor or a combination of modalities such as the camera, the lidar, the radar, etc. While cameras are limited to bad weather conditions, the radar is robust and provides additional information about

the objects' characteristics (as the distance and the velocity) which can reduce fatal errors. Thus, the combination of various sensor types can make a significant enhancement.

Datasets are decisive for researchers and developers as most of the tools and algorithms have to be tested and then trained before functioning on the road. A typical practice consists in testing and validating the developed algorithms on annotated datasets.

In this context, various autonomous driving datasets have been published in order to enhance reasarch for environment perception such as Kitti [55], Kaist Multi-Spectral [30], nuScene [20], etc. Most of these datasets are multimodal, combining different heterogeneous modalities.
While some of the existing datasets use the narrow-band radar, UWB carries richer information. The UWB radar provides a signal that results from the reflection of a transmitted UWB pulse on the object. The deformation of the initial wave represents the signature of the object. This signature contains information about the distance, the material and the shape of the object. In fact, the UWB radar offers a huge interest in short-range applications, but its employment addresses a main challenge which distinguishes the target over noise and static clutter.

From another point of view, different objects have distinguishable acoustic signatures that may help recognize each of them. In spite of the usefulness of the acoustic data, we notice that none of the-state-of-the-art ITSs benchmarks use the acoustic modality.

Moreover, nowadays, the employment of deep learning remains unavoidable when developing object detection systems due to the achieved success in this area of research. Therefore, in order to increase accuracy and provide tangible improvements, we have to cope with their complexity. Further, training such algorithms needs huge amounts of data. For these reasons and motivated by the fact that current databases lack in radar and acoustic sensors adoption, the development of new ones is required.

1.3 CONTRIBUTIONS

In this thesis, we propose a new multi-object detection framework by exploring either uni-modality or multi-modality via developing a multi-modal dataset for advanced urban environment perception.

In order to achieve the thesis aim, the contributions of this work are as follows:

1. The first contribution consists in developing **OLIMP** (A HeterOgeneous MuLtimodal Dataset for Advanced EnvIronMent Perception) with over than $47,000$ samples. It is a new heterogeneous dataset collected using a camera, a UWB radar, a narrow-band radar and a microphone. To the best of our knowledge, this is the first dataset that includes UWB signals and acoustic data including several challenges related to dense urban traffic. Thus, potential advancement could be accomplished in the environment perception area using our dataset.

2. Our second contribution consists in proposing a new fusion framework that combines data acquired from the different sensors used in the introduced dataset in order to achieve better performances for the obstacle detection task. This fusion framework demonstrates the usefulness of the introduced dataset and it mainly highlights the importance of multimodality in environment perception.

3. A segmentation approach to detect ROIs in an environment perception-dedicated UWB radar is our third contribution. Specifically, we implement a differential analysis of the entropy of UWB signals to detect ROIs. We evaluate our technique on OLIMP. The obtained results show higher performance in terms of obstacle detection compared to the implemented state-of-the-art techniques, as well as stable robustness even with low amplitude signals.

4. The fourth contribution is a novel framework that exploits RNNs with UWB signals for the detection of multiple road obstacles. We evaluate our approach on the OLIMP dataset. The obtained results show that the system outperforms the other implemented related techniques in terms of obstacle detection by learning the temporal relationship between the data sequences.

1.4 THESIS OUTLINE

This document is organized as follows.

In chapter 2, the state of the art is presented by firstly putting the topic of object detection and environment perception into context with the autonomous driving. In addition, the exploited sensors and the relative main challenges will be detailed. Afterwards, we review road object detectors using uni-modal and multi-modal systems by explaining the possible fusion levels.

Chapter 3 presents the developed dataset. This chapter goes further into the implementation details in terms of employed sensors, synchronisation setup, annotation process and relative challenges. In addition to this, a fusion framework exploiting different fusion levels is presented to highlight the potential enhancement that can be achieved using the introduced dataset.

The following chapter goes deeper into investigating the UWB radar. In chapter 4, two UWB-based detectors are proposed. The first detector aims to segment obstacles within UWB signals via an entropy-based approach. Regarding the second framework, it is a deep-based detector that takes advantage of the LSTM network to distinguish noise from real targets. A comparative study with the state-of-the-art techniques is conducted, and the obtained results are discussed in this chapter.

Finally, chapter 5 presents the conclusion of this thesis and some perspectives.

2

ENVIRONMENT PERCEPTION SYSTEM: STATE OF THE ART

Contents

2.1 INTRODUCTION

In this chapter, we firstly present the state of the art relative to environment perception systems including road object detection for intelligent and AVs. The relevant sensors and the fusion challenges are described. Afterwards, a detailed overview of object detection methods is exposed.

This chapter is organized as follows. Section 2 presents an environment perception background that includes the Advanced Driver Assistance System (ADAS) applications, the employed sensors and the challenges related to the vehicle's surroundings perception task. Section 3 exhibits a review of object detectors from one uni-modal system to multi-modal systems for road obstacles. The performance evaluation metrics of the detectors are detailed in Section 4. Finally, section 5 refers to the conclusion of this chapter.

2.2 ENVIRONMENT PERCEPTION BACKGROUND

According to the WHO, every day, around 3,700 people are killed in road traffic crashes and over than half of those are vulnerable road users: pedestrians, cyclists, motorcyclists, drivers [139]. Existing surveys point out that the human errors are one of the principal causes of road accidents. These errors can be distinguished in fatigue, heedlessness and immature behaviors. For this reason, it becomes critical to equip vehicles with safety systems to provide security to drivers and vulnerable road users.

Aside from these mentioned consequences, road accidents have additional undesirable side effects. They are identified as a considerable cause of energy consumption and air pollution [97]. Furthermore, people waste countless hours in the urban traffic environment.

Addressing the aforementioned issues of safety, efficiency and pollution remains a primordial concern to guarantee a better life quality. Therefore, the development of intelligent vehicles is a viable solution to the mentioned problems in order to ensure security by avoiding accidents and providing an economic, comfortable and intelligent driving.

Recently, AVs receive worldwide attention thanks to the considerable advancement and progress that have been achieved in this field of research. These improvements are made on account of the prompt advances constructed in information, electronics and

communications technologies, and by employing essentially AI. In fact, an AV is a car that moves safely and takes decisions by sensing its surroundings with little or no human intervention in real traffic conditions.

Some basic characterizations for automation levels are currently set according to the Society of Automotive Engineers (SAE) [193] for better understanding of self-driving cars. Different levels of automation are defined and they cover degrees from no automation to complete control. Otherwise, the higher level is the more the vehicle's monitoring responsibilities.

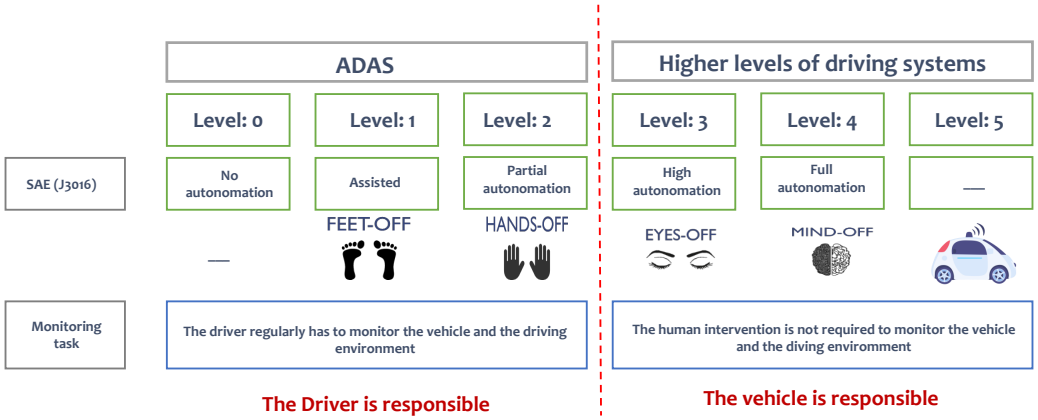


Figure 2.1: Levels of driving automation according to SAE J3016 [89]

According to the SAE international's J3016, there exist 6 levels describing the state of automation for a vehicle [89]. These levels are presented as below and described in figure 2.1.

- Level 0: It is labeled as no automation and the driver entirely controls the vehicle continually.
- Level 1: It is known as 'feet off'. The automated system and the driver all together control the car. Automatic parking and the automatic breaking are given as examples. Besides, the driver must be always prepared to retake total control any time.
- Level 2: It is entitled 'hands off' and the automated system is responsible for accelerating, breaking and steering the car. The driver monitors the driving and must be ready to rapidly intervene if the system fails to respond correctly.

- Level 3: It is called ‘eyes off’. At this stage, the driver can ignore safe driving task, like texting, working and even watching a movie. The car handles situations that require an instantaneous response such as the emergency breaking. However, the driver still must be willing to intervene in limited time fixed by the manufacturer.
- Level 4: It is known as ‘minds off’. At this level, the driver cedes the full control to the vehicle and is not expected to control it at any time. For example, the driver can safely sleep.
- Level 5: It is entitled ‘steering wheel optional’, and the human intervention is not required (e.g. a robotic taxi).

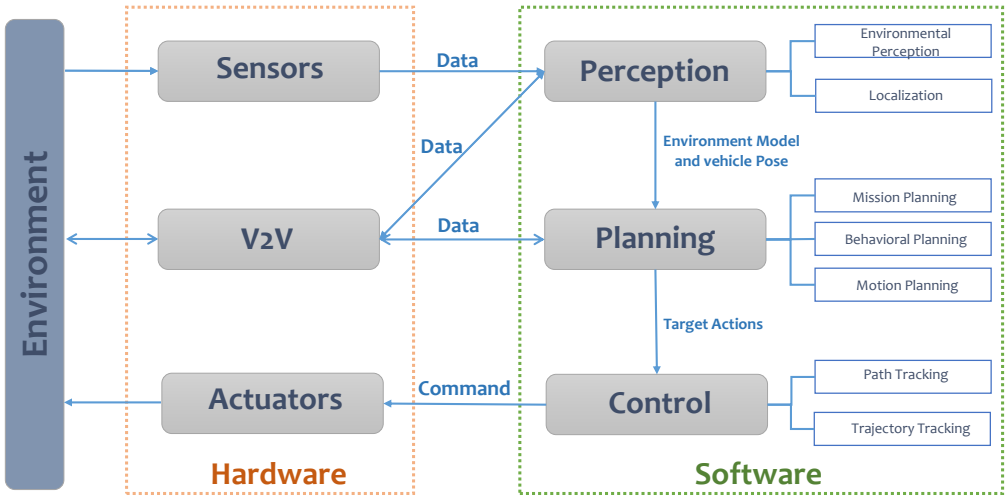


Figure 2.2: Typical autonomous vehicle framework

A typical autonomous driving framework can be categorized into three main stages: **perception, planning and control.** The mentioned stages and the car’s interactions with the environment are depicted in Figure 2.2.

The perception process is responsible of accurately perceiving the vehicle’s surroundings through suitable sensors. Environmental perception aims to understand the environment, by positioning the obstacles, detecting the road signs, and categorizing the acquired data. Localization refers to the determination of the vehicle’s position compared to the environment.
The planning process makes decisions in order to achieve autonomous driving goals, such as avoiding obstacles.

Finally, *the control* stage refers to the execution of the planned actions generated by the planning process.

2.2.1 *Advanced driver-assistance systems: ADAS*

The supporting and assessment of the driver in critical conditions are provided by ADAS applications. Literally, these applications present extra traffic information, an overview of the driver's behavior, and environment information to guarantee an efficient performance.

In fact, the advancement of driver assistance systems was set out at the end of the 70's through an Anti-lock Braking System (ABS) embedded into a serial production [192]. Afterwards, the improvement steps in this area can be distributed to three classes: network sensors, proprioceptive and exteroceptive sensors. As regards to the proprioceptive sensors, they are responsible for detecting danger situations and then responding through analyzing the vehicle's behavior. Concerning exteroceptive sensors, like the radar, ultrasonic, lidar and vision sensors, they are capable of responding on a prior stage and predicting dangers [86]. The employment of these sensors ensures ADAS applications. Some of these ADAS applications are detailed below and illustrated in figure 2.3.

- Blind spot detection: It observes the vehicle's adjacent area. Thus, it warns the driver of the existence of obstacles in the blind spot, via a visual sign in the side view mirror or through an audible signal.
- Road cross traffic alert : It helps avoiding accidents when the driver reverses out of parking. This functionality provides a visible signal and an audible warning if an object is detected in the driving direction's reverse.
- Traffic sign assist : It automatically recognizes traffic signs, even the signs of other countries. Therefore, safer and relaxed driving is provided.
- Lane Departure Warning : It is in charge of scanning the edges of the road and detecting when the car is deviating the lane of the road. Thus, the driver will be warned via an hyptical signal as a steering wheel vibration or through a visual warning.
- Emergency Brake Assist : It is an active braking support that automatically brakes on critical situations, so, rear-end collisions will be refrained absolutely and pedestrians will be protected.

- Adaptive Cruise Control : It it is able to control the distance from the car ahead. Moreover, it alerts the driver or slows the vehicle's speed if the respective distance becomes inadequate (too small).

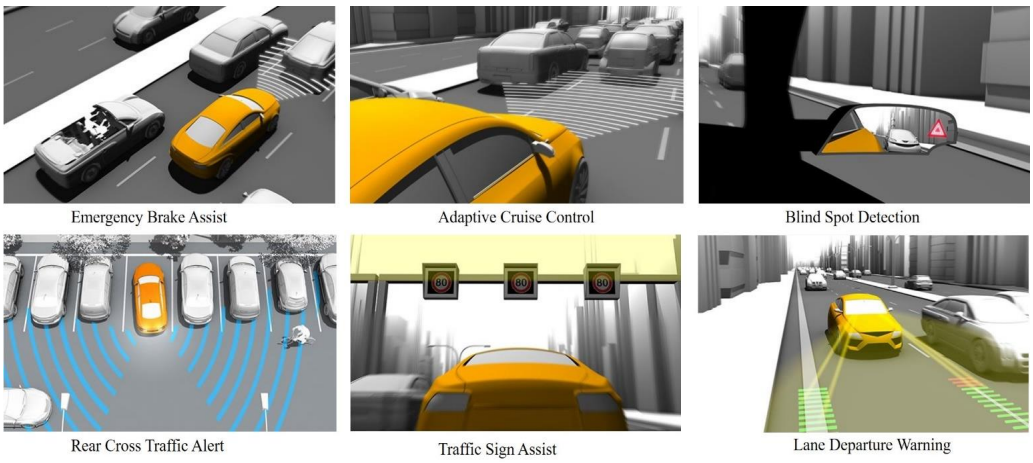


Figure 2.3: ADAS

2.2.2 *Sensing Modalities for intelligent transportation systems*

Sensors collect information about road conditions, and vehicle's surrounding can be categorized into two classes: active and passive. Active sensors diffuse signals, and based on the reflected signal it identifies targets, such as the radar and the lidar. Passive sensors acquire data without diffusing, and cameras are the most widely used example.

- ***Radar***

  Nowadays, the radar is well exploited in many fields such as mapping, meteorology and especially in the area of automation [120]. In fact, the principal goal of a radar system is to detect the existence of one or more targets in the area of interest. A radar simultaneously transmits and receives electromagnetic waves in frequency bands between 3 MHz and 300 GHz, and it extracts information (range, position, velocity) using the reflected EM waves from the targets [120]. It is robust against fog, rain, bad weather and lightning conditions (day and night). Furthermore, the automotive radar systems can be divided into three classes: short range radars that are mostly employed for parking assist, medium range radars used for rear collision avoidance, and long range radar utilized for adaptive cruise control [103].

- ***Lidar***

  Light detection and ranging known as lidar, is considered as one of the dominant technologies in the field of AVs. It is fixed on the roof of the car and it spins regularly. Lidar is a laser scanner that provides a 360-degrees of visibility of the vehicle's environment and measures the range from 1m to 60m depending on the sensor. Besides, it illuminates the target with a pulsed laser light and measures the reflected pulses through a sensor [8]. The sensor outputs point cloud data that include the position (x, y, z coordinates) of the objects and their intensity information which indicates the object reflectively. There are three major lidar types used in the autonomous driving field: 2D lidar, 3D lidar and solid state lidar. This sensor is expensive and sensible to weather conditions such as fog and snow owing to the diffraction of light within these circumstances.

- ***Camera***

  Cameras are considered as the eyes of the vehicle through outfitting the car with several cameras at all angles. In fact, two versions of visible cameras are used in this field which are mono and stereo vision [192]. Fora mono vision camera, it is usually employed to understand the essential surroundings as detecting the speed limit signs or lane marking. Concerning stereo vision cameras, they are analogous to human eyes and provide two video sources. In fact, the utilization of such a technology helps the vehicle to understand the traffic flow and obstacles' positions. In addition to mono and stereo vision, there exists a night vision system, which adopts infrared cameras. Although cameras are sensible to lighting variations and weather conditions, they are the only sensors capable of detecting color, texture and contrast information [140].

- ***Ultrasonic***

  It is a low-cost sensor that sends sound waves in high frequency in order to determinate the object's distance. Actually, it is widely used in this field to detect near obstacles; however, it is affected by noise and interference [135].

The aforementioned sensors are highlighted in figure 2.4 to highlight the benefits and limits of each sensor, Table 2.1 summarized the mentioned sensors' characteristics.

Sensor type	**Avantages**	**Disadvantages**	**Max working distance (meters)**
Radar	- Long working distance - Provide distance, velocity without any computing resources - Robust in all weather conditions (fog, rain, dust) and during nighttime	- Generate false alarms easily - Sensitive to interference	5m-250m
Camera	- Excellent discreniblity - Extensive information in images (color, texture, etc.) - High resolution and larger FOV	- Weather susceptible - Light interfernce - Heavy calculation	Depending on lens
Lidar	- High range and angle resolution - Wide FOV	- Very expensive - Insufferable for bad weather - Power consuming	150 m
Ultrasonic	- Inexpensive	- Low resolution - Sensitive to noise and interference - Short range detection	4m

Table 2.1: ADAS sensors' characteristics

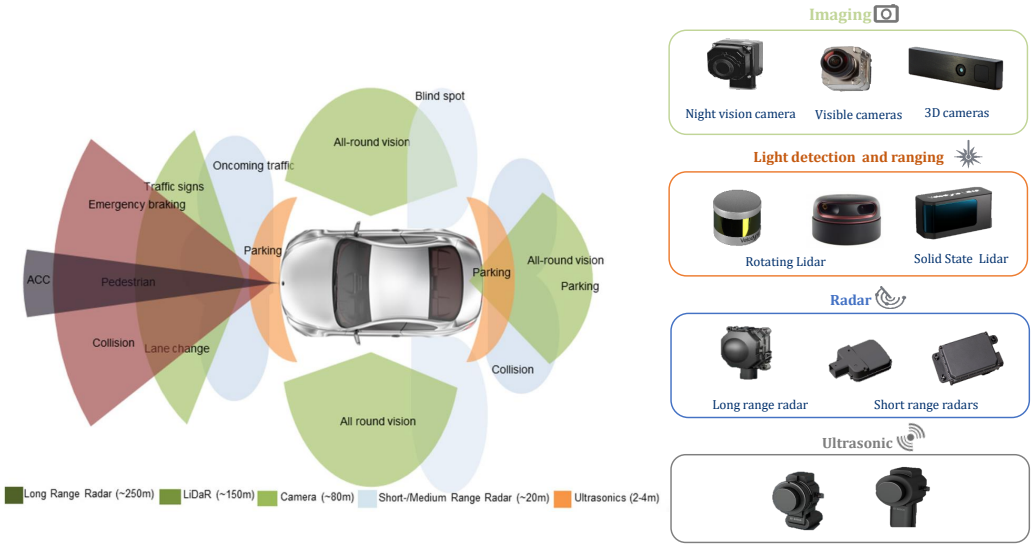


Figure 2.4: Sensors employed for ITSs

2.2.3 *Multi-modal environment perception challenges*

The environment perception stage is the first and most primordial process for automated driving. It provides the vehicle with decisive information on the driving environment over time. Thus, the vehicle should determine its position in order to correctly interrupt the data acquired from the perceptional sensors. This task is known as localization. Afterwards, the vehicle senses its surroundings through the employed sensor. Robust detection and classification of stationary and moving obstacles is essential to correctly perceive the environment. Furthermore, the surrounding objects' positions and velocities are determined and tracked during this stage, and even future states can be predicted.

Thereby, the development of a reliable perception system remains a challenging task, because the car must perceive its surroundings in real world situations that include uncontrolled and complex scenarios such as urban environment. Figure 2.5 illustrates a complex urban scenario using multi-modal sensors, and including multiple relevant traffic participants.

Accordingly, the current challenges related to the environment perception are caused by the complex outdoor environments that include different road agents, as well as the presented requirement to develop efficient algorithms for real-time perception. In addition, variable lighting and adverse weather conditions, uncontrolled backgrounds

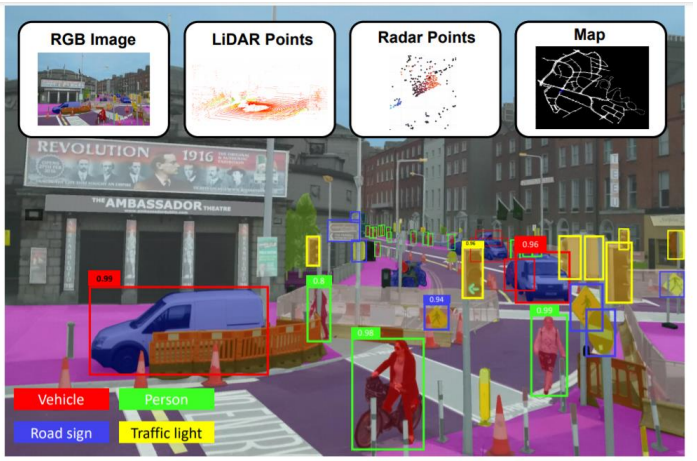


Figure 2.5: Complex urban environment for autonomous driving [47]

and the occlusion of multiple objects remain critical challenges for the perception process for intelligent vehicles.

On account of the aforementioned challenges and as a small error can engender fatal accidents, the environment perception system should be:

(i) ***accurate***: It must to provide precise information about the driving environment.

(ii) ***robust***: It should work accurately in adverse weather, under various situations and circumstances, and even when the sensor degraded.

(iii) ***real-time***: It should have a real-time response particularly when driving at a high speed.

To achieve achieving these goals, the environment perception system takes advantage of multi-modal sensors to completely perceive the vehicle surroundings. As already indicated, the most utilized sensors in the automation field are: cameras and lidar and radar sensors. In fact, every sensor has its typical advantages and disadvantages, as mentioned in section 2.2.2 so each of these sensors can be employed in different situations. Thus, a comparative study of the most used sensors according to the autonomous requirements is presented in figure 2.6. As depicted from the figure, it becomes obvious that all the requirements for autonomous driving can not be ensured by an individual sensor type. However, a combination of two sensors or more can achieve good results for the environment perception task. Hence, the fusion of various sensing modalities permits exploiting their complementary properties.

	Camera	Radar	Lidar	Autonomous requirement
Object localization task	Medium	High	High	High
Object Classification task	High	Medium	Low	High
Distance estimation	Low	High	High	High
Range of visiblity	Medium	High	Medium	High
Functionality in bad weather	Low	High	Low	High
Functionality in poor lighting	Low	High	Low	High
Cost	Medium	Medium	High	High
Amount of data	Medium	High	Medium	High

Figure 2.6: Comparison of perception environment sensors

2.3 OBSTACLE DETECTION

As mentioned in Chapter 1, in our work we focus particularly on one of the fundamental environment perception dilemmas which is object detection. Object detection points out the identification of the objects of interest locations and determines their sizes. This section discusses the state of the art of object detection and especially the work related to object detection for ADAS applications.

2.3.1 *Recent developments in object detection*

The state of the art of object detection can be practically divided to pre and post-deep learning appearance. In fact, preceding the progress of deep learning approaches, the literature were principally based on hand crafted features such as Histogram of Oriented Gradients (HOG) or the scale-invariant feature transform (SIFT), which are employed with traditional classification methods like the Support Vector Machine (SVM), Adaboost, etc. Some major pre-deep learning contributions made in the object detection field are listed in the following.

- Cascade of weak classifiers : It is one of the basic approaches that was proposed by Viola and Jones [167]. Haar features are extracted, and Adaboost with cascade classifiers are used for object detection. The proposed technique is based mainly on the sliding-window principle.

- HOG : In [39], HOG features were introduced by Dalal and Triggs, and they are based on the edge directions within the image. SVM classifier was utilized for object classification.

- Deformable Parts Model (DPM) : It is a graphical model proposed by Felzenszwalb et al. Actually, the DPM introduced to face objects' deformation in the image [46]. DPM is based on the fact that each object is composed of its parts.

- Selective Search (SS) : This method was put forward by Uijlings et al. [165]. The SS technique generates independent object proposals. In fact, the input image is segmented into different scales, which generates several ROIs. A similarity comparison based on the color, size, texture, etc. is made to merge the redundant regions and set will be suggested afterwards.

Nowadays, with the significant advancement of employing deep learning, a relevant impact on the filed of object detection is made. It is considered a powerful tool as it can learn hierarchical features for large amounts of data. Correspondingly, several methods have been proposed to tackle the object detection dilemma. The state of the art related to deep object detection can be divided in two categories: the two-stage or one-stage pipelines. The two-stage network is also known as the region-based technique and the one-stage is called the single shot object detector.

- ***Region-based object detection network:***

  In the primary stage, several ROIs are extracted and are considered as object candidates. Afterwards, these region candidates are confirmed and classified. Then, classification scores and locations are refined. The pioneering work that utilized deep learning for object detection was OverFeat [152] and R-CNNs (Region-based ConvNetworks)[57]. Indeed, ROIs are firstly generated using the sliding window technique for OverFeat and selective search for R-CNNs. Subsequently, the suggested ROIs are processed via CNNs to extract relevant features for the classification and the regression of the bounding boxes. For the SPPnet (Spatial Pyramid Pooling networks) [63] and Fast-RCNN [56] networks, the features are directly generated from feature maps by employing a larger CNNs on the entire image (e.g. Resnet [64], VGG [155], GoogleLeNet [159]). In Faster R-CNNs [134], the object detection pipeline is unified and a region proposal network is introduced for generating region proposals. Following this line, in [37] the R-FCN (Region-based Fully Convolutional Networks) was put forward, which was a fully-convolutional network. In fact, the fully-connected layers of the RPN are replaced

with convolutional layers. Run time and accuracy are increased for object detection performances.

- ***Single shot object detection network:***

  This type of model proposes to straightly match the feature maps with the bounding boxes and classification scores via a unified CNN network. For example, YOLO (You Only Look Once) [132] models the object detection task as a regression dilemma where the object proposal generation process is eliminated. The bounding boxes are directly regressed from the CNN. The SSD (Single Shot Detection) [93] treats objects with different sizes via regressing numerous feature maps of various resolutions. Small convolutional filters are used for predicting multi-scale bounding boxes.

Generally, the region-based object detectors like the Faster-RCNN tend to reach higher detection accuracy owing to the generation of various region proposals and the refinement model. This comes with a higher execution time, high power computation and a more complicated training process. In contrast, the one-stage networks are faster and can be optimized easier, which makes them well-suited for real-time applications. Yet, in terms of accuracy, they are under-performed in comparison to the two-stage object detectors [70]. Figure 2.7 shows the performance evolution for basic CNN-based architectures from 2012 to 2019 evaluated on the ImageNet 2012 dataset. This comparison is based on the achieved accuracy and the number of parameters illustrated with a varied circle. The mixed colors in the figure indicate that two architecture kinds are combined [69].

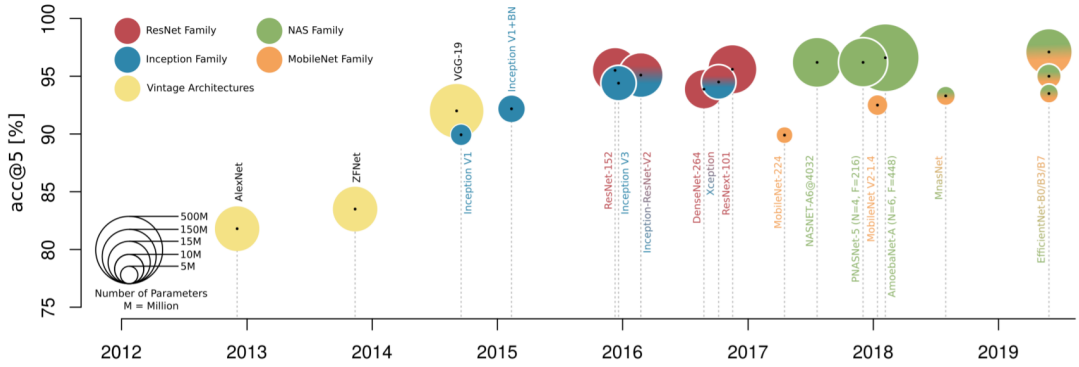


Figure 2.7: Performance evaluation for basic CNN-based architecture from 2012 to 2019 carried on ImageNet 2012 dataset [69]

2.3.2 *Object detection in ADAS*

Dynamic environments contain various static and moving obstacles that interact with each other. Therefore, distinguishing these objects is an essential task of the perception environment stage. In fact, the development of object detection systems for autonomous driving is specifically challenging as it has high requirements in terms of real-time, accuracy, and robustness performances. The results obtained from object detection are commonly transferred to the other units such as decision making. Thus, a reliable object detection system is a prerequisite for safe driving under complex and uncontrolled driving environments [173].

Object detection in ADAS faces various challenges that are related to the complex driving environment, which includes a cluttered background and various road-agents. These agents have different types (traffic signs, pedestrians, cars, motorcyclists, etc) with various sizes. Moreover, many obstacles can be occluded, so occlusion adds additional challenges to the object detection task. Likewise, bad lighting and weather conditions still affect the detection performances significantly.

Furthermore, when developing object detection systems, it is essential to consider some critical aspects. For the input data, the questions that remain are: Are there any available multi-modal or uni modal datasets? Are the data of high quality and labeled?. Furthermore, several important questions should be answered: Which modalities should be used or combined? How are required data represented and how can we process them correctly?. Accordingly, which fusion methods can be applied and at which stage can we have an accurate and reliable object detection system? Correspondingly, various challenges related to object detection when developing ADAS should be considered .

2.3.3 *Uni-modal based systems for road obstacle detection*

Complex driving situations often present various obstacles. Some works has focused on 2D detection, while some others deal with 3D object detection, which includes more challenges thanks to the development of complex datasets. To address this challenge, the use of a single modality or a combination of various ones has been adopted in the literature. In this section, we summarize various existing techniques for uni-modal based systems for obstacle detection.

2.3.3.1 *Vision sensor-based object detection*

One of the essential sensors used for observing the vehicle's surroundings is the camera, so, computer-vision based approaches have been widely employed for obstacle detection. Some research has focused on using only one camera while others have employed several cameras to obtain multi views of the vehicle surroundings and accurately detect the obstacles.

Existing vision-based methods can be distinguished into three categories: stereo-vision, classical-based and deep learning based methods. The stereo-based methods use two cameras that generate a depth map due to their capacity of 3D scene reconsecration. Accordingly, the objects are segmented within the depth map relative to their spatial locations. These approaches are able to detect various obstacles with different shapes and various motion statuses. Moreover, they can accurately determine the distance and the 3D geometric size [184].

Classic visual obstacle detection approaches employ hand-crafted descriptors such as the HOG [39], Aggregated Channel Features (ACF) [41] or the Integral Channel Features (ICF) [42]. A pedestrian detector was proposed in [130], and features were extracted from numerous image scales using the ACF. The approach consisted in training several models using multiple resolution. Finally, the bounding boxes generated by each model were concatenated. In fact, an improvement of 6 % was achieved in terms of average precision over the employment of single resolution. A vehicle proposal location framework was introduced in [189]. The suggested algorithm was a graph-based method that generates accurate region vehicle proposals compared with traditional approaches. The experiments were carried out on the PASCAL VOC2007 and the Kitti dataset. In [94], obstacle regions were segmented via a specific threshold to obtain binary images. Yet, this technique depended on the defined threshold, so some obstacles were not well segmented and noise also affected the obstacle detection process. The V-disparity algorithm has been used in various studies for obstacle detection [25] [54] [109]. Nevertheless, this approach is sensitive to large objects that will influence the detection of small objects. The latest classical object detector is the DPM which has achieved significant improvements in the object detection area. Yet, its detection accuracy is still limited for driving object detection and its computation complexity remains very high. In [185], HOG features were incorporated with disparity maps via a modified DPM. The disparity maps were determined from the stereo images using the semi-global matching method. In [170], a traffic sign detection system was put forward, which included three processes: image preprocessing, detection and recognition. An RGB color segmentation was adopted and followed by a shape matching technique. Then, SVM was employed as a classifier. A

vehicle detection system that relied on a stereo vision classifier was proposed in [112]. The fusion Haar, the Local Binary Patterns (LBP) and the HOG features provided good detection results. In [62], an ROI obstacle was defined based on the distinctive characteristics of the optical flow of the background over the optical flow of the obstacle region.
Despite the fact that promising results have been found using the classical techniques for object detection, these methods tend to fail in numerous complex driving environments that comprise different objects' sizes and types, and critical challenges.

While classical object detectors get stuck in the bottleneck, there are a wide deep learning models that boost research for visual object detection. In [178], a novel RPN was suggested which exploited subcategory information to lead the region proposal process. The input of the RPN was a pyramid image where it was processed via various convolution and pooling layers to generate for every scale one feature map. Afterwards, a convolutional layer was added for the subcategory object detection, where every filter would fit with a specific subcategory. Correspondingly, the ROI generation layer would use a thresholding technique on the heat maps to provide the object candidates. Cai et al. [22] proposed a multi-scale CNN (MS-CNN) in order to handle the wide variation of the objects' size. The MS-CNN performed the object detection over various scales of feature layers, that lead to an improved detection rates with a speed of 15 frames per second. The authors in [183] employed scale dependent pooling to exploit suitable convolutional features related to the sacle of the object candidate in order to improve the detection accuracy. Furthermore, cascaded rejection classifiers utilized the convolutional features and excluded the negative object candidates using the cascaded manner to speed up the detection. The combination of the two contributions achieved better detection accuracy on Kitti, PASCAL and Inner-city datasets. A Recurrent Rolling Convolution (RRC) network was proposed in [133], inspired by the SSD architecture. The pedestrian and cyclist detection accuracy achieved higher performances than the state of the art on the Kitti dataset. Nevertheless, the RRC was a complex model with a high computation time. In [173], visual object detection was guaranteed using three enhancements for the CNN. They include a deconvolution and a fusion of the CNN feature maps to obtain deeper features. Moreover, a soft non maximal suppression was adopted to address the occlusion challenge. The Kitti dataset was exploited in this work. Zhang et al. proposed in [190] a cascaded R-CNN to extract pyramids that represent the weighted multi-scale features using the dot product and the softmax to enhance the traffic sign accuracy detection. The authors in [131] used the mask R-CNN to detect the object and the optical flow method was utilized to analyze the detected object movements.

2.3.3.2 *Lidar sensor-based object detection*

Compared to vision based detection, reliable depth information is provided by a lidar point cloud. The lidar point cloud is a set of points represented in space and corresponds to a 3D object or shape. Moreover, 3D voxels can be generated via grouping the neighbor points onto sets, so the computational cost is reduced. These data can be exploited to accurately detect 3D objects and distinguish their shapes. For this reason, lidar-based object detection has a deceive role for the visual perception system of ADAS to tackle the camera-related limitations.

The object detection using lidar point clouds faces some challenges such as the sparsity of the lidar points, the high variability of the point density, the occlusion, the corresponding pose variation and the non-uniform sampling related to the 3D space [78].

The existing works on lidar-aided object detection methods can be distinguished into three categories [174]:

(i) Projection-based methods: Several studies have focused on the projection of the lidar points onto the bird's view images. Then, 3D bounding boxes are regressed according to the extracted features from these images [9] [176]

(ii) Point-voxel methods: Lidar data are represented as 3D voxels, and 3D CNNs are applied for the prediction of 3D bounding boxes [98] [181]

(iii) Pointnets-based methods: Lidar points are processed directly through neural models without a pre-processing step [87] [158].

In the following, we briefly review some existing work that has focused on object detection from the point cloud. Classical methods using lidar data for object detection have used clustering algorithms in order to segment the point clouds and assign the obtained groups to multiple classes [166]. The Vote3Deep modal was introduced by Engelcke et al. [44] based on sparse convolutional layers. L1 regularisation was adopted in this work for adequate processing of the 3D lidar data. Other work has exploited some handcrafted features associated to the spatial relations between the segmented points as fast-point-feature histogram [160] and the pin images [50] were extracted to identify the categories of clusters. PointNet [126] and PointNet++ [127] were introduced for processing the point sets and achieved good results in indoor environments. Nonetheless, with the expanding amount of point cloud data involved in developing 3D object detectors, computational power and huge memory requirements increased. In [107], the point cloud was represented under five views: the distance, the size, the reflection, the range

and the height. These views were the CNN inputs. The projection of the 3D data onto 2D views generated the loss of important information that was critical for the detection task, particularly in congested scenes. An efficient deep-learning-based network, entitled VoxelNet, was put forward in [191]. Features were directly extracted from sparse points within the 3D voxel grid, so, remarkable results were achieved on the Kitti dataset. However, the detection accuracy rate decreased slightly as the distance between the voxels in the grid and was smaller than the distance of 3D lidar data. In addition, various voxel-based methods have been proposed such as SECOND [181] and PointPillar [88].

Although lidar-based approaches can achieve remarkable object detection performances, these methods require high resolution and a precise and available lidar point cloud. Furthermore, the lidar has typically blind areas when detecting objects closer to the vehicle.

2.3.3.3 *Radar sensor-based object detection*

Radar sensors provide rich information about the vehicle's surroundings on account of the received signals and the characteristics extracted from these signals such as: the amplitude, the range, the velocity, the Doppler spectrum, etc. The acquired radar data depend on the sensor type and its characteristics. In fact, it can be depicted by 2D feature maps. Next, it can be processed via a CNN or a point cloud, or it can be represented as 1 dimensional (1D) signals that can be treated, afterwards, to detect the obstacles within these signals.

Various object detection methods firstly cluster the radar targets according to their characteristics (e.g. range, velocity, azimuth) to a group of object proposals. Next, these clusters are classified. In [124], DBSCAN [45] was used to cluster the radar targets. Subsequently, various cluster-wise features were extracted as the variance of the velocity. The performance of numerous classifiers was compared for pedestrian detection as the SVM and the Random Forest (RF). The authors in [150] also used DBSCAN to cluster different targets. Then the LSTM and the RF were compared for the multi-class (pedestrian, car, truck, etc.) detection task.

Wile clustering-based approaches are widely employed, it is often observed that objects can be wrongly merged or can be split apart. The performance depends essentially on the initial clustering step that relies on the definition of the suitable parameters that must be utilized for all classes (as the radius of the cluster using DBSCAN). Furthermore, small objects may be ignored using such methods.

To tackle these challenges, some researchers have suggested to classify each target separately rather than in clusters. In [149], inspired by the results achieved using then lidar point cloud, the authors proposed to process radar data via PointNet++ [127] to detect dynamic objects. Yet, a single radar frame datum is too sparse.

On the other hand, others have chosen to extract features from the acquired radar data for classification. The Radar Cross-Section (RCS) values were obtained using different frequencies, and the angles of simulated and measured obstacles have constituted the feature vectors. Consequently, SVM was used for the classification of indoor targets [19]. The received amplitude, the Doppler spectrum and the range characteristics were extracted from the radar received signals. An SVM was employed to classify pedestrians and cars [65]. A similar approach was suggested in [66].

Despite the fact that deep learning techniques have achieved remarkable advancement in various fields of research, only few studies have employed radar data with deep learning methods. In fact, to apply such networks, radar data require a pre-prossecing step where the radar reflections are represented as an image-like data [149]. For static object detection, occupancy grids were used as inputs for CNNs in [95] [96]. Kim et al. [83] opted for utilizing convolutional Recurrent Neural Networks (RNNs) for moving object classification via range-velocity maps. The range-velocity images are obtained by transforming the time-series radar data by adopting the 2D discrete fourier transform. The obtained results show that the LSTM-based network is able to learn the dynamics of the lateral movements related to the vulnerable road users in the time-series radar images. In [3], the spectrogram extracted from the time-frequency signals was represented as 2D images. These images fed the stacked auto-encoders for extracting high-level radar features. The authors in [84] transformed the acquired radar data to range-angle representation. Following this, YOlO was trained on the transformed data. The system accuracy reached 90% on a self-recorded dataset.

Table 2.2 summarizes the shortcomings of the aforementioned methods of object detection according to their different characteristics.

Generally, it is challenging to utilize only data acquired from a single sensor in a complex environment, specially for critical applications. The reasons can be caused by the sensor's shortages, the sensed environment, or both. Perceptional sensors suffer from various limitations and inadequacies, which can degrade the detection performance. Apart from the imperfections of sensors, the challenging environment conditions have a huge impact on their outputs as weather circumstances (eg. rain, fog) or illumination conditions (e.g. nighttime, low light). Besides, each sensor is suitable for a specific use case, scenario and application. For instance, the radar provides certain information about

Detection mode	Detection method	Detection characteristics
Computer-vision	*Binary threshold*	Cannot completely segment obstacles
	UV-disparity	Sensitive to large obstacles and smal obsacles can be ignored
	CNN	Large amount of data required Time-consuming
	Optical flow	Complex implementation and time-consuming
	Multi-camera	Suffer from fisheye distortion Camera calbiration required
Lidar	-	Sparse data
Radar	-	Sparse data

Table 2.2: Characteristics of different object detection methods

the obstacle's distance, but its sparse data require a constructive processing step, otherwise objects cannot be detected. Concerning the vision-based system, it can accurately detect objects as pedestrians, but it is time-consuming. Additionally, using lidar data can guarantee impressive performance, but this sensor is still easily affected by rain, fog and dust.

Analyzing the advantages and limitations of employing each sensor separately, we observe that the sensors mentioned above are complementary for the object detection task. In other words, each sensor compensates the limits of the other sensor. Therefore, based on these reasons, the need for employing multiple sensors and merging the acquired data remain essential to enhance the efficiency of environment perception tasks.

2.3.4 *Multi-modal based systems for road obstacle detection*

Since incomplete and unreliable information can result in fatal driving situations due to the challenging driving environment, combining data collected from disparate sensors remains a good solution to upgrade the system's overall detection performance.
Even though merging information obtained from numerous sensing modalities is a challenging task, thanks to advanced sensor technologies and the progress in data processing algorithms, with the improvements in hardware, the fusion is becoming realizable.

In this section, an overview of the adopted fusion strategies in the autonomous driving field is presented and some multi-modal perception systems for the obstacle detection task are reviewed.

2.3.4.1 *Data fusion methods for automotive applications*

The combination of data acquired from various modalities is of great interest. In this regard, most existing work fuses RGB images with lidar point clouds. In addition, some further work couples RGB images with thermal ones. However, we highlight that there is a lack of research on combining radar data with images.

In fact, for object detection, there exist two predominant categories when fusing various modalities: hand-crafted feature methods and deep-neural-network approaches. Furthermore, the fusion of sensing modalities can be achieved at three possible stages: early, intermediate or late. These levels are detailed in figure 2.8. For simplicity, the sensing modalities are restricted to two.

- ***Low level***

  The low-level fusion is also known as an early or signal level. During this level, the raw data acquired from multiple sensors are directly combined in order to obtain merged data that can be used for successive tasks. For instance, the lidar depth map is combined with the color camera data to define an RGB-D format that is processed afterwards.
  This level has a low memory budget and low computation requirements. However, it is sensitive to data misalignment among the employed sensors, which can be caused by different sampling rates or calibration errors.

- ***Intermediate level***

  The intermediate level or the medium-level fusion intends to extract features from several data collected via multiple sensors. These features are, then, combined into a feature vector that serves as the input for the subsequent process. An example is the extraction of features from RGB images as HOG features and from the depth map independently. These features are concatenated to a single feature vector.
  Although this fusion type enables the system to learn different feature representations at several depths, it is not straightforward to identify an optimal way to couple them for a specific architecture.

- ***Late level***

  The late or high-level fusion is as known as decision level. The obtained decisions that process each datum from different sources separately are combined to define the final decision. For instance, combining the detected bounding boxes obtained from the object detection algorithm processed on the RGB and depth map separately determines the final detections using the voting method, for example.
  The late fusion level is highly flexible; e.g. when an additional sensing modality is introduced, it does not affect the overall architecture. Nevertheless, it suffers from high memory and computation cost requirements.

In addition, there exists multi-level fusion (known as the hybrid-level). This level focuses on the integration of various data at different levels of abstraction.

Moreover, based on the literature, five fusion operations are mainly used to fuse multiple modalities based on a deep architecture [48]: 1) Addition, 2) Average mean, 3) Concatenation, 4) Ensemble: It is used to combine the ROIs for object detection, 5) Mixture of Experts: This operation tends to modal explicitly the weights of the feature maps.

It should be pointed out that there is no evidence that confirms that one fusion level is better than the others. The performances are extremely dependent on the data, the employed modalities and the network.

In the following, some multi-modal perception systems that exploit fusion are reviewed. A summary of these approaches is presented in Table 2.3.

2.3.4.2 *Fusion using camera and lidar data*

Multiple studies have proved that fusing images with lidar data improves the accuracy of the object detection process, particularly for far ranges and small obstacles [4]. There are three techniques to combine lidar point clouds with camera images. Firstly, the results obtained from training images and lidar points separately are merged. Secondly, the targets are detected using camera images. Afterwards, the confirmation of the results is provided using the lidar point clouds. Finally, the third method consists in defining ROIs utilizing lidar data, and the camera is used to detect the objects.

Conzalez et al. [58] used transformed depth maps and RGB images as inputs to detect pedestrians. In this work, the objects' poses in multi view were taken into account, and

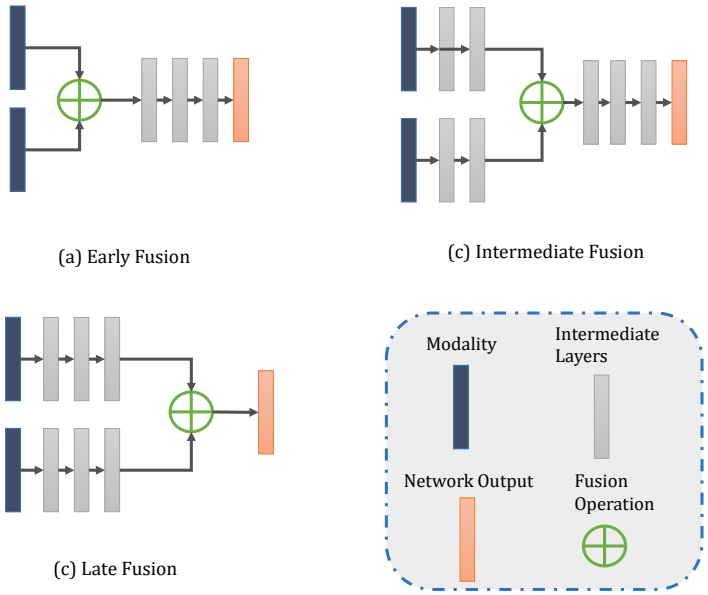


Figure 2.8: Data fusion levels: (a) early fusion, (b) intermediate fusion, and (c) late fusion

intermediate fusion and late level fusion were used. For the intermediate stage, they fused features extracted from HOG and LBP descriptors, with the SVM classifier. With regard to high-level fusion, they coupled decisions obtained from the training of a detector on each modality. In this case, the feature level fusion guaranteed a better performance.
In [156], a point fusion method was proposed where lidar points were mapped onto the image plane and features were extracted from the image using a pre-trained 2D detector. Afterwards, features were concatenated via a VoxelNet architecture. The authors in [11] suggested an architecture based on two single-stage detectors. The information provided by lidar data (height, distance, intensity) was transformed into images. These latter with RGB images, were the inputs. These data were processed by VGG16 [155] to provide features. Afterwards, an SSD [93] network was adapted to generate the bounding boxes of 2D cars in foggy weather based on a deep feature exchange that relied principally on feature concatenation. In the work of Xu et al. [179], the raw data acquired from lidar were proceeded by a PointNet architecture, and images features were extracted via the CNN. The obtained results were then pooled in order to locate the coordinates of the 3D bounding boxes. Qi et al. [125] adapted a similar approach in their work. In the work of [115], object proposals were generated using a segmentation method applied on the data of lidar point clouds and RGB images. After that, the candidates generated from lidar data and images trained two separate CNNs in order to classify the proposals. The output decisions were combined using a basic belief assignment to associate the

bounding boxes. Then a CNN model was implemented to determine the final decision along with the SVM.

2.3.4.3 *Fusion using camera and thermal camera data*

Even though visual cameras are affected by weather and lighting conditions, thermal cameras are robust to nighttime and daytime circumstances because they detect the object's heat reflected by the infrared radiation. For this reason, the combination of the provided data can ensure detailed scene understanding as they are correlated in terms of illumination conditions.
Hwang et al. [72] introduced an extension of the ACF dedicated for pedestrian detection. The extended modal consisted of a multispectral ACF obtained from the augmentation of the thermal intensity and the employment of HOG features as additional ones. In [80], visible and thermal images were fused according to two approaches in the intention to detect people. The first consisted in encoding the two types of images using independent encoders, and the encoded features are merged and then decoded back in order to generate a single fused image that would be the input of a Residential Network (ResNet) architecture. This technique was called DenseFuse. The second proposed method was an intermediate level fusion technique. Indeed, ResNet-152 was employed separately for infrared and visual images. Thereafter, the extracted features were concatenated into a single array that would serve as the input of the fully connected layer.
Early fusion and late fusion based on the CNN architecture to couple infrared and visible images were investigated in [169]. The early fusion method consisted in combining the pixels captured from the two modalities, in contrast to late fusion where two sub-networks would provide a feature representation for both modalities. These representations were fused on a supplementary fully connected layer. Besides, the proposals were generated using the ACF+THOG detector. According to the obtained results, a pre-trained late fusion method evaluated on KAIST multispectral dataset guaranteed better performances. In [90], an illumination-aware architecture was proposed based on the Faster R-CNN [134]. Infrared and visible images were respectively the inputs of two separate sub-networks. Meanwhile, an illumination aware network was developed to estimate an illumination value from color images. Thereafter, an illumination weight layer is integrated in order to determine the fusion weights for both modalities. Consequently, the final decision was achieved by weighting the final results obtained from the two sub-networks due to the estimated fusion weights.

2.3.4.4 *Fusion using radar and camera data*

For obstacle detection, the radar and the camera are two complementary sensors, but only a few studies have addressed this challenge. Similar to the other kinds of sensing combinations, the three types of fusion can be applied to couple these modalities.
In [172], radar tracks generated the ROIs in the images. Afterwards, for the vision module, a symmetry algorithm and a contour detection technique were applied to the ROIs to identify vehicles. The goal of the work presented in [17] was to detect pedestrians. The radar sensor provided a list of tracks and the ACF object detector was adopted to generate a list of identified pedestrians in the images. Subsequently, the fusion of the obtained descions was ensured using the Dempster Shafter method. Wang et al. [171] proposed a decision approach to fuse radar data and images. The YOLO [132] network is employed in this work to detect vehicles from visible images. The radar sensor detects the centroid of the obstacles. Afterwards, these detections were projected on the image plane. Finally, the results obtained from the two modalities are combined. A real-time Radar Region Proposals network (RPNP) was developed in [110]. The suggested network consisted in generating ROIs based only on radar detection. In fact, the tracks are mapped into images so that anchor boxes are proposed, which are inspired by fast R-CNN architecture. Then, these boxes are scaled according to the distance of the objects to have accurate detection. Radar data are transformed into images in [23] in order to be combined with RGB images. Actually, these data will be proceeded via the ResNet network separately. Accordingly, features are concatenated after the second block of ResNet.

To fuse different modalities for understanding the vehicle surroundings, many approaches have employed deep neural network architectures, while others are based on hand crafted features. From the aforementioned reviewed studies, we observe that the fusion performance depends mainly on the sensing modalities, the quality of data and the selected architecture. For fusion operations, feature concatenation is a widely exploited method, specifically in early and intermediate levels. Likewise, the addition and mixture of experts are mainly employed for intermediate and high stages.

Ref	Object class	Sensing modality processing	Hand crafted features	Network pipeline	Fusion level	Used dataset
			Lidar and camera fusion			
[58]	Pedestrian	-Depth maps generated from lidar sensor -RGB images	-HOG -LBP	-	Intermediate Late	Kitti
[156]	3D car	-Lidar voxel -RGB images processed via 2D image detector	-	Early Intermediate	Kitti	
[11]	2D cars (foggy weather)	-Depth, intensity and height information acquired from lidar and processed via VGG16 -RGB images processed via VGG16	-	Early and intermediate layers	Self-recorded dataset	
[179]	Car Pedestrian Cyclist	-Lidar raw data processed by PointNet -RGB image features extracted via CNN	-	CNN	Early	Kitti SUN-RGBD
			Infrared and visible camera fusion			
[80]	Persons	-RGB images and thermal images encoded for early fusion	-	ResNet-152	Early Intermediate	KAIST-Multispectral
[169]	Pedestrian	-RGB images and thermal images processed via CaffeNet	-	R-CNN	Early Late	KAIST Pedestrian dataset
[60]	Pedestrian	-RGB images and thermal images processed via VGG16	-	Faster R-CNN	Early Intermediate Late	KAIST Pedestrian dataset
			Radar and camera fusion			
[172]	Vehicles	-Tracks from radar sensor -RGB images	-Symmetry detection algoritm -Active contour detection	-	Intermediate	Real-workd recorded dataset
[17]	Pedestrian	-Radar generates list of tracks -RGB images	ACF object detector	-	Late	Real-workd recorded dataset
[171]	Vehicles	-Detections from radar -RGB images	-	YOlO	Late	Self-recorded dataset under rainny weather
[110]	Car, Person, Motorcycle, Truck,Bicycle and Bus	-Tracks from rear radars -RGB images from the rear camera	-	Fast R-CNN	Early	Two substes from nuScenes dataset
[23]	2D Vehicle	-Radar range proceeded by ResNet -RGB images proceeded by ResNet.	-	One stage detector	Intermediate	Self recorded dataset
[114]	Car,Bus, Motorcyle, Truck,Trailer, Bicyle,Human	-Radar data transformed to image plane -RGB images	-	VGG	low	nuScences TMU Self recorded dataset

Table 2.3: Summary of fusion approaches for obstacle detection

2.4 PERFORMANCE EVALUATION METRICS FOR INTELLIGENT TRANSPORTATION SYSTEMS

Commonly, object detector outcomes include a list of detected bounding boxes, confidence levels and affected classes. In fact, the bounding box is mostly presented by its bottom-right and top-left coordinates ($x_{initial}$,$y_{initial}$,x_{end},y_{end}), with an exception for the YOlO algorithm.
The most common performance metrics used to evaluate the object detector are the Average Precision (AP) and and its derivatives as the mean Average Precision (mAP) over all the objects' classes. Previously, some concepts should be reviewed which are shared among these metrics. The basic ones are presented below :

- True Positive (**TP**): It is the accurate detection of a ground truth bounding box.
- False Positive (**FP**): It is the incorrect detection related to a misplaced bounding box of an existing object or a nonexistant object.
- False Negative (**FN**): It is a miss-detected ground-truth bounding box.

It is important to point that a True Negative (TN) result is not considered in the scope of object detection, since there are limitless bounding boxes that should not be detected among any image.

The aforementioned definitions require distinguishing correct detection from incorrect detection. For this reason, the Intersection Over Union (IOU) measure is used. This metric is based on the Jaccard index, which represents a measurement of similarity between two sets of data [73]. In the object detection context, the IOU is employed to measure the overlapping area among the ground-truth bounding box (B_{Gt} presented by the orange color) and the predicted bounding box (B_P presented by the green color) divided by the area of the union between them [117]. The IOU is presented by Equation 2.1 and illustrated in figure 2.9.

$$IOU = J(B_P, B_{Gt}) = \frac{B_P \bigcap B_{Gt}}{B_P \bigcup B_{Gt}} \tag{2.1}$$

Subsequently, the obtained IOU value is compared to a defined threshold (**thr**) to decide if the detection is accurate or inaccurate. If the IOU $\geq$ **thr** , then the detection is correct. Otherwise, if the IOU $<$ **thr**, then the detection is incorrect.

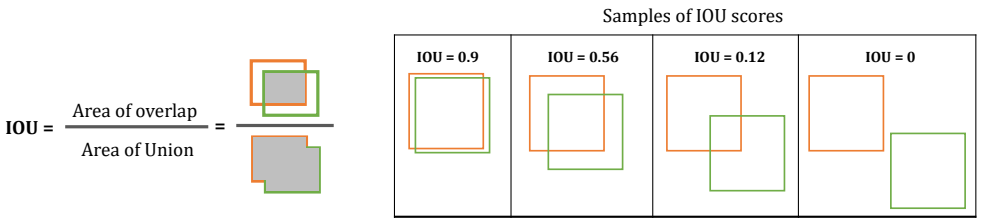


Figure 2.9: Intersection-over-union (IOU) metric samples of single object.

After determining the TP, FP and FN through the IOU calculation, the assessment of the detectors rely mostly on the Precision (P), the Recall (R) and the F1-score metrics. The P presents the ability of the detector to identify just the relevant objects. This metric introduces the percentage of the correct positive predictions. Concerning the R, it presents the ability of the detector to identify all the relevant samples, i.e. all the ground-truth bounding boxes. It exhibits the percentage of the accurate positive predictions within all the considered ground truths. The F1-score is the metric that balances P and R. The metrics mentioned above are presented respectively by Equation 2.2, Equation 2.3 and Equation 2.4.

$$P = \frac{\text{TP}}{\text{TP + FP}} = \frac{\text{TP}}{\text{All detection}} \tag{2.2}$$

$$R = \frac{\text{TP}}{\text{TP+ FP}} = \frac{\text{TP}}{\text{All ground truths}} \tag{2.3}$$

$$F1 - score = 2 \times \frac{\text{P . R}}{\text{P+R}} \tag{2.4}$$

Besides, an object detector achieves a good performance when it identifies all the ground-truth objects, so FN is equal to 0 (high recall) while finding only relevant obstacles so FP is equal 0 (high precision). Hence, the detector is considered good when its precision rate is high while its recall rate increases, which means if the confidence threshold is changed, the P and R will remain high. Thus, the AP metric is calculated via Equation 2.5. It uses the all-point interpolation concept where the precision is interpolate

at each level and it takes the maximum precision value whose recall rate is greater than R_n.

$$AP = \sum_n \left(R_n - R_{n-1}\right) P_n \tag{2.5}$$

The mAP is the metric that measures the accuracy of the object detector over all the classes considered in a specific dataset. The mAP is the average of th AP over all classes and it is presented by Equation 2.6.

$$mAP = \frac{1}{N} \sum_i^N AP_i \tag{2.6}$$

where AP_i is the AP of the ith class and N is the total number of the classes that are being evaluated.

2.5 CONCLUSION

In this chapter, we have presented the state-of-the-art of object detection dilemma for environment perception for intelligent and autonomous systems. By analysing the advantages and the disadvantages of using each sensor separately, we have concluded that these sensors are complementary to perceive the vehicle's surroundings. Following, we reviewed the fusion strategies and object detectors based on combing different sensors at various levels.

In the next chapter, we will review the datasets used in this field of research and we will present our multimodal developed dataset for advanced environment perception. After that, a fusion framework will be presented.

3

OLIMP: A HETEROGENEOUS MULTIMODAL DATASET FOR ADVANCED ENVIRONMENT PERCEPTION

Contents

3.1 INTRODUCTION

As we mentioned in the previous chapter, almost of the multi-modal environment perception approaches are based on supervised learning. Accordingly, multi-modal datasets including labeled ground-truth is needed in order to train such methods, so, to develop road object detectors.
This chapter summarizes numerous published real-world datasets regarding employed sensors, the recording conditions, their size and labels. Consequently, we detail our proposed multi-modal dataset and a new fusion framework that combines data acquired from the different sensors used in our dataset to achieve better performances for obstacle detection task.

Section 3.2 refers to an exhaustive overview of the available public environment perception databases. The proposed dataset is introduced in section 3.3 which includes the sensors setup, the scenarios, the challenges and the dataset details. The section 3.4 exhibits the proposed fusion framework and the obtained results.

3.2 EXISTING PUBLIC MULTIMODAL ENVIRONMENT PERCEPTION DATABASES

Public multimodal datasets are indispensable for autonomous driving's advancement. In the last decade, several datasets have been released for this purpose, Kitti [55] dataset and Cityscapes [31] dataset are considered the first datasets that have addressed real-world challenges. Until few years ago, datasets that contain only sparsely annotated data were satisfactory to treat several problems. But nowadays, with the evolution of deep learning techniques the exploitation of such datasets is insufficient [79].

In fact, the training of deep models requires datasets with a huge number of labeled data though collecting such amount of data is not an obvious task. Hence, this requirement has led to the development of several new sophisticated autonomous driving datasets [61]. In this section, we review various existing public monomodal and multimodal environment perception databases by detailing and observing the characteristics of each one. Table 3.2 exhibit an overview of various environment perception datasets.

- **Kitti** : It is a vision benchmark dataset that was released in 2012 and comprises stereo camera, Velodyne lidar and inertial sensors [55]. Within the introduction of this database, various vision tasks were launched as pedestrian detection, road

detection, etc. It was recorded in six different emplacement with cluttered scenes and it provides over 200k boxes that was manually labeled. Nevertheless, only 3D objects that exist in frontal view are annotated and it covers only daytime conditions. Moreover, the preeminent limitation of Kitti database is the small amount of data that is not suitable for deep learning algorithms.

- **CamVid** : The University of Cambridge has introduced a new driving dataset named CamVid. It was the first that contains videos with semantic segmentation labels related to 32 classes. However, the size of this dataset is small; it contains a limited number of sequences: only four scenes [18].

- **Cityscapes** : It is a dataset that was published in 2016 [31]. It covers urban traffic scenarios in 50 cities, in only spring and summer and it includes 30 categories. Cityscapes consists of a pixel-level and instance-level segmentation labeling. Indeed, it contains mainly images and few videos with 5000 images which have fine-annotation over 20000 images along coarse annotations.

- **BBD100k**: It was recorded in 2016 in four different regions in the US [188]. It is considered as the largest driving video dataset due to its diversity in terms of data and driving conditions. The database comprises 100k videos containing almost 1000 hours recorded under different weather conditions. Indeed, only one image is selected from each video sequence for labelling likewise Cityscapes dataset. Ten thousands images are labeled in pixel level and bounding box labels are provided for 100k images.

- **Kaist Multi-Spectral**: It is a multimodal database that was repeatedly collected in urban, residential and campus environments [30]. Several sensors were fixed on the vehicle, namely: stereo camera, thermal camera, GNSS, 3D lidar and inertial sensors. Moreover, it covers a diverse time slots (day, night, morning, sunset, etc.) and the annotation is provided in 2D. Compared to the newest released datasets, the size of the Kaist dataset is limited.

- **ApolloScape** : Compared to Kitti and Cityscape databases, the ApolloScape dataset [71] contains an extensive amount of data and has many properties that will be detailed in the following. In fact, it includes stereo driving sequences that reach over one hundred hours of recording under diverse day times and about 144k images. It covers also 2D and 3D pixel-level segmentation, instance segmentation, lane marking and depth. Further, in the intention to label such a database, the authors developed several tools customized mainly for the annotation process. However, data acquired from lidar is used to provide only static depth maps.

- **H3D** : It was introduced in 2019 , and it considers various complex and congested scenes over 160 [119]. Three cameras, a lidar, a GPS and inertial sensors were used to collect this dataset. The main challenge addressed in this dataset is 3D multi-object detection and tracking. In fact, it consists of 1.1M 3D boxes annotated data, which includes over 27 k frames. Plus, objects are labeled in 360° view. Eight classes was taken into consideration when recording this dataset : car, pedestrian, cyclist, truck, misc, animals, motorcyclist and bus. It is true that the dataset comprises rich scenes and annotation with a particular size, nevertheless, the data was registered under daytime conditions.

- **BLVD**: The dataset introduced in [180] and entitled BLVD does not focus on static obstacle detection only, but especially on dynamic object detection. Indeed, this dataset proposes a platform that involves 4D tracking (3D+temporal), 5D interactive recognition events and 5D intention prediction. It includes 3 categories : vehicle, pedestrian and rider, the data was recorded in daytime and nighttime conditions. It provides 120k frames with a 5D semantic annotation and beyond 249 3D annotation.

- **nuScene (nuTonomy scenes)**: It is the first dataset that involves the three preeminent sensors exploited to ensure an autonomous driving which are a lidar, 5 radars and 6 cameras [20]. This database consists of 1000 scenes where the duration of each scene is 20s. The annotation provided is a 3D bounding boxes specified for 23 categories. The data were gathered under several lighting and weather conditions : daytime, nighttime and rain . This new released dataset is rich in terms of utilised sensors, size, acquisition conditions diversity, amount of data with 1.4M frames and annotation numbers. Yet, the main issue of this dataset is the class imbalance represented by the inequality number of examples of infrequent and ordinary object classes.

- **A2D2**: The A2D2 (Audi Autonomous Driving Dataset) is recorded via six cameras and five Lidar in order to provide a full 360° coverage. It includes 41,277 frames along semantic segmentation images and point cloud labels. In addition, this dataset is the only dataset that contains vehicle bus as the steering wheel angle, the throttle, and the braking. The A2D2 data were recorded on cities, highways and country roads in the south of Germany under sunny, cloudy and rainy weather conditions.

Other than the autonomous driving databases mentioned previously, it exists additional datasets that are released for the same purpose, such as the Oxford Robotcar [100], Udacity [164] and DBNet Dataset [26].

Dataset	Autonomous Driving Task						
	Multi-Object detection	Object tracking	Optical flow	Lane detection	Semantic segmentation	SLAM	3D vision
CamVid [18]	x				x		x
kitti [55]	x	x	x	x	x	x	x
Cityscapes [31]	x				x		
BDD100k [188]	x	x		x	x		
Kaist Multispectral [30]	x	x		x		x	
ApolloScape [71]					x		
H3D [119]	x	x					
BLVD [180]		x					
nuScenes [20]	x						

Table 3.1: Categorization of some autonomous driving datasets by task

Furthermore, some multi-modal virtual datasets or virtual simulators have been developed in order to generate variable driving situations, specifically the dangerous scenarios that can't be or hard to be collected in real-world. For instance, Gaidon et al. [51] develop a virtual Kitti database, in [137] and in [177] virtual dataset have been generated based on game engines as GTA-V. Dosovitskiy et al. [43] build an open-source simulator for simulating multiple sensors. Even though various virtual dataset are available now, the question that remains is how accurate the simulator can represent real-world circumstances.

As the databases are mainly released to enhance the scenes understanding and environment perception, we provide in Table 3.1 a categorization of the most important autonomous driving datasets according to a particular tasks.

From table 3.1, we can notice that most of the reviewed datasets were dedicated to multi-object detection as it is an inevitable process in the autonomous world. Likewise, there are favorable number of datasets dedicated to object tracking, lane recognition and semantic recognition, but in return, just a few ones can be used to optical flow exploitation and SLAM (Simultaneous localization and mapping) process.

Following the exhibition of the most datsets, we provide above a comparative study in terms of the recording conditions, the employed sensors, the dataset size, etc.

- ***Sensing Modalities :*** In terms of sensing modalities, all the examined datasets contain RGB images acquired from one or more cameras or video in HEVC (high efficiency video coding) standard or in recent coding [49]. Lidar sensor also have been well exploited. For radar data, it is only presented in nuScenes dataset [20] and the newly released Oxford Radar RobotCar Dataset [7] and Astyx HiRes2019 [105]. It is a very limited number despite that this

sensor provide rich information and helps in the environment perception and taking the right decisions. For that reason, nowadays, it becomes essential to exploit radar sensor in developing autonomous driving datasets.

– ***Recording Conditions*** : The majority of the collected data is specialized in urban driving, and was recorded in different locations: Europe, the United States, Asian cities, etc. this variance in locations allows us to have a global view of roads conditions, environments, etc.

  One of the important criteria to have a complete dataset, is that it is collected under different lighting and weather conditions in order to cover various scenarios [76]. Nonetheless Kitti dataset is broadly used in this field of research, the variety of its recording environmental conditions is reduced: it is gathered just under daytime and sunlit days, similar to CamVid, CityScapes and H3D datasets. In order to enrich light recording conditions [30],[20], [188],[71] and [180] collected data considering both daytime and nighttime all day long. Concerning the diversity of weather conditions, only BDD100k and nuScenes covers rain and snow situations.
  Actually, seasonal changes are not well covered as the majority of the databases were recorded in short periods.

– ***Dataset Size*** : We notice that since 2016, the number of the published datasets becomes extensive because of this importance in the development of self-driving cars. As the dataset size plays a key role in this field, it varies from 1,569 frames to above than 11 million frames, and it has grown over the years. nuScences is considered as the largest dataset with 1,4M frames. Yet, compared to the size of the image datasets related to the computer vision community, the environment perception datasets remain relatively small.

– ***Labels*** : Depending on the principal aim of the published dataset, objects are labeled into various categories. Comparing the object classes existing in each dataset, we can observe that the number of examples attributed to each class is imbalanced. For example, we compare the samples related to two different classes: car and pedestrian for nuScenes, Kitti and Kaist Multispectral databases. We can observe that there are much more car labels than pedestrian labels, as shown in figure 3.1.

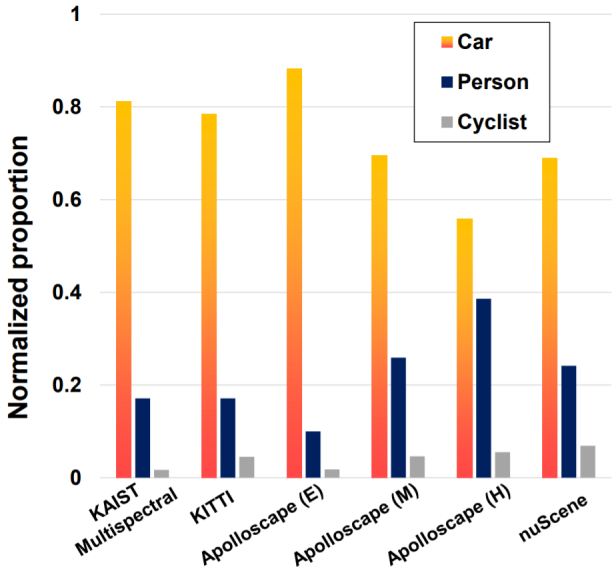


Figure 3.1: Comparison a normalized percentage of objects' samples: car, person and cyclist related to Kitti, KAIST Multispectral, Apolloscape (E: easy, M: moderate and H: hard) and nuScene dataset [48].

Dataset	Year	Modalities	Size	Annotation		Varity					Categories	Recording Cities
				2D	3D	Daytime	Nighttime	Fog	Rain	Snow		
CamVid [18]	2008	Camera	4 scenes	x		x					32 classes	Cambridge
Kitti [55]	2012	Camera Lidar Inertial sensors	22 scenes	x	x	x					8 classes	Karlsruhe
Cityscapes [31]	2016	Camera	-	x		x					30 classes	50 cities
BDD100k [188]	2017	Camera	100k	x		x	x		x	x	10 classes	Four regions in US
Kaist Multispectral [30]	2018	Camera(stereo) Thermal camera 3D lidar GNSS Inertial sensors	-	x		x	x					Seoul
ApolloScape [71]	2018	3 camera(stereo) 3D lidar GNSS Inertial sensors	-	x	x	x	x				35 classes	Four regions in China
H3D [119]	2019	3 cameras Lidar GPS Inertial sensors	160		x	x					8 classes	San Francisco
BLVD [180]	2019	3 cameras Lidar GPS Inertial sensors			x	x	x				3 classes	Changshu
nuScenes [20]	2019	6 cameras Lidar 5 radars	1000	x	x	x	x	x	x		23 classes	Boston & Singapore

Table 3.2: Overview of some autonomous driving datasets

3.3 PROPOSED DATASET

As mentioned in the previous section, the importance of multimodal perception techniques for ITS and the extent of research efforts in this direction emphasize the need for multimodal datasets that explore complementary sensors. Therefore, in this section we present our proposed dataset and we detail its implementation, challenges and opportunities.

3.3.1 *Background*

Various autonomous driving datasets have been published in order to enhance research for environment perception. Most of these datasets are multimodal, combining different heterogeneous modalities. While some of the existing datasets use narrow-band radar, the UWB radar carries richer information. The UWB radar provides a signal that results from the reflection of a transmitted UWB pulse on the object. The deformation of the initial wave represents the signature of the object. This signature contains information about the distance, the material and the shape of the object. Moreover, different objects have distinguishable acoustic signatures that may help recognize each of them. In spite of the usefulness of the acoustic data, we notice that none of the state of the art ITS benchmarks uses acoustic modality.

Thus, we introduce OLIMP (A heterOgeneous MuLtimodal Dataset for Advanced EnvIronMent Perception) (https://sites.google.com/view/ihsen-alouani), a new public dataset for road environment perception. The introduced OLIMP dataset is a multimodal synchronized dataset that was collected using four heterogeneous sensors to better understand the vehicle's environment.

Our benchmark contains four complementary modalities namely: UWB radar data, narrow-band streams, images and acoustic data. In fact, camera is affected by degraded condition such as foggy weather, while, UWB radar is not influenced by neither luminosity nor weather conditions. The acoustic data is orthogonal to the vision field. Concerning the narrow-band radar, it provides position and velocity. To the best of our knowledge, OLIMP is the first benchmark that contains UWB radar data and acoustic data.

The data was collected in the campus of the Polytechnic University Hauts-de-France in Valenciennes - France (Valenciennes is known for its foggy weather). Data was captured during 3 months and consists of ***47354*** synchronized frames.

3.3.2 *Hardware and data acquisition*

On the one hand, we used four heterogeneous sensors: a monocular camera, an UWB radar which is a short range radar, a narrow-band radar that is a long range radar and a microphone. On the other hand, we exploited the EFFIBOX platform to acquire data from the different sensors simultaneously [161]. In table 3.3, we highlight the sensors' characteristics and their technical details.

- ☐ UMAIN radar: it is an UWB radar. The exploited kit is called HST-D3 developed by the UMAIN corporation [163]. The kit comprises a UWB short radar with a Rasperby Pi 3 for the acquisition. Following, the received radar raw data are transmitted to the computer through the Raspberry Pi that is connected via TCP/IP protocol. More details about the UMAIN radar are provided in the next chapter.

- ☐ Narrow-band radar (ARS 404-X): This Premium sensor from Continental is a long range radar that is able to detect multiple obstacles up to 250 meters. It genertaes raw data that include: distance, velocity and radar cross section RCS [29]. Data are transmitted to the EFFIBOX platform via CAN bus.

- ☐ The EFFIBOX platform : is a software developed in (C/C++) dedicated to the design of multi-sensor embedded applications. In addition, diverse adequate development functionalities are available such as : acquiring and saving sensor streams, processing/post-processing, visualization, etc.

It should be pointed that, the EFFOBOX platform has its own API(Application Programming Interface) to communicate with the ARS radar, the network camera and the microphone in order to acquire and record data. For the UMAIN radar, we developed our API so that the EFFIBOX can communicate with the radar. Then, the acquired data has been decoded following a particular protocol provided by the company. Besides, the frame acquired from the ARS radar are decoded also according to the protocol provided by the Continental datasheet.

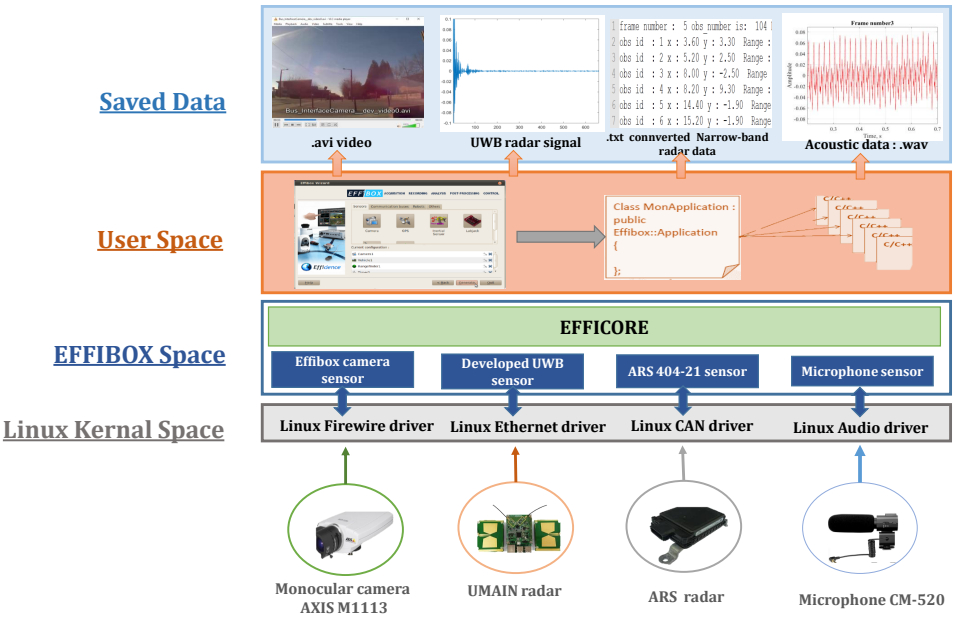


Figure 3.2: Data acquisition architecture

3.3.3 *Sensors embedding*

With regard to the sensors configuration, we designed a structure where all the sensors are placed in the front view. To simplify the data fusion, the narrow-band and the UWB radars and the camera were mounted on the same vertical axis. Figure 3.2 shows the proposed data acquisition architecture, and figure 3.3 highlights the structure setup.

3.3.4 *Sensor synchronisation*

To develop an efficient autonomous driving dataset, sensor synchronisation is a challenging and inevitable task. In fact, sensor fusion often requires that collected data from all the sensors have to be captured on the same time as each sensor has its own latency. To illustrate this phenomena, figure 3.4 (a) shows a samples of a simultaneous acquisition from different data streams.

Therefore, we developed our method to achieve an accurate alignment between the modalities' data streams. In the simultaneous data recording process, we register times-

Sensor	Sepcification	Measure latency
AXISM1113 [5]	-A monocular camera, RGB images, 25 FPS, 640x480 resolution, angle of view: 65°-25°, 50Hz	20 ms
UMAIN radar [163]	-≤ to 6 range, provides signals, each obstacle has its own signature, 4GHz.	22,5 ms
ARS 404-X [29]	-≤ to 250 range, provides distance; velocity and RCS, 77GHz, $\pm 0.40\ m$ accuracy for far range, $\pm 0.10\ m$ accuracy for near range	72 ms
Microphone CM-520	-≤ to 20 range, +10dB sensitivity, it fits well with video cameras, 50Hz-16Khz for frequency response	Not Applicable

Table 3.3: Sensor specifications and properties. Measure latency is the time necessary to collect one complete data stream from the sensor.

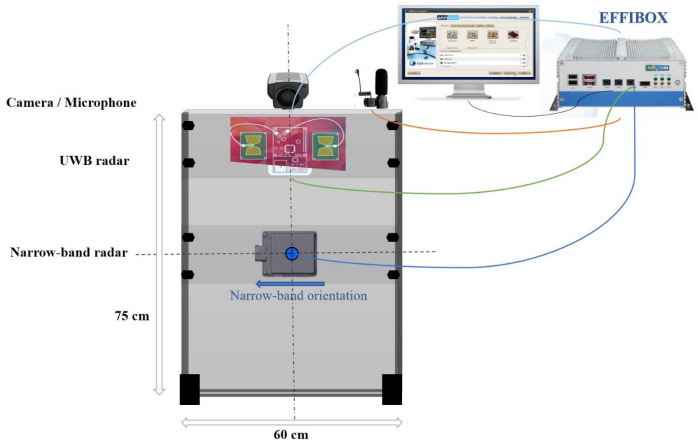


Figure 3.3: Structure setup

tamps relative to each sensor separately. We first start with synchronizing the radars and the camera. Since these sensors have different frequencies and time responses, we choose the narrow-band radar as a primary sensor. This is explained by the fact that the narrow-band radar is the slowest among these sensors; it has the highest latency of a complete measure compared to the other modalities as shown in Table 3.3. In fact, the narrow-band radar raw data is represented in the form of a stream of discrete measures. Each one of these measures comprises a main data frame including the obstacle's number followed by successive information about each obstacle (distance, velocity, dynamic property, etc.). Once a narrow-band measure is taken, we capture its timestamp and look for the camera frame as well as the UWB frame that have the closest timestamp to the synchronization narrow-band timestamp.

Regarding the acoustic modality, the frame corresponds to an analog signal (sound). The challenge is to find the most suitable time window size that: (i) corresponds to the exact scene recorded at given timestamp, and (ii) is long enough to hold meaningful information about the scene. After thorough explorations, we empirically choose an optimal window size of 5 seconds for acoustic signal frame. This frame is recorded according to the narrow-band synchronization timestamp mentioned above.

Overall, the proposed algorithm consists of selecting the timestamp acquisition of every narrow-band measure and find the corresponding frames of the other sensors which have the closest timestamp. The frames synchronization step is illustrated in figure 3.4 (b).

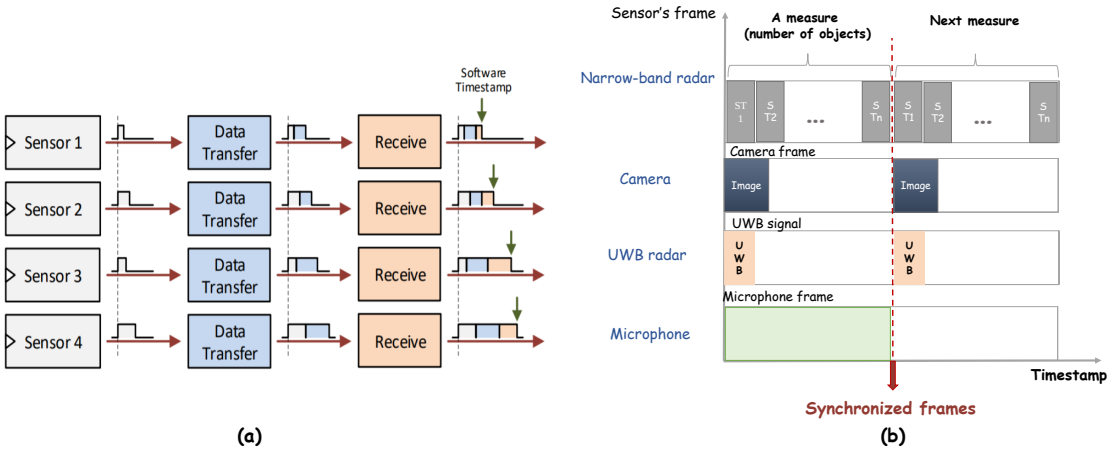


Figure 3.4: Frames synchronisation (ST: object stream, n: number of obstacles in the scene).

3.3.5 *Labeling process*

In addition to the background, we consider four classes: pedestrian, cyclist, vehicle and tram since these are the most probably encountered possibilities in an urban transport environment. The vehicle class contains cars, tracks, etc.
For the labeling process, we manually annotate data consecutively one image per three as this task is time consuming and the changes between two successive images are practically negligible. We avoided automatic annotation to have a high quality labeled ground truth. Thus, we used the Matlab Image Labeler toolbox whom we have the license as semi-automatic labeler tool.

Data annotation includes 2D bounding boxes that present respectively x, y, the width W and the height H of the object in pixels.

3.3.6 *Scenario selection and data formats*

In order to collect raw sensor data, we carefully choose diverse driving situations. The duration of scenes differs and depends mainly on the situation complexity.
While recording our dataset, we consider diverse challenges that will be detailed in the following subsection. Accordingly, we emphasize the data variety through employing different locations (8 emplacements) that vary in terms of structures, environment, road markings, traffic signs, etc. Some of these recording emplacements are illustrated in figure 3.5.
Driving situations are carefully selected and collected under different lighting conditions, we covered also sunny, cloudy and snowy weather.
For data format, the dataset provides synchronized frames of each scenario, the data are stored as : RGB images, .txt files presenting UWB radar signals, .txt files of narrow-band radar data stream and .wav microphone files.

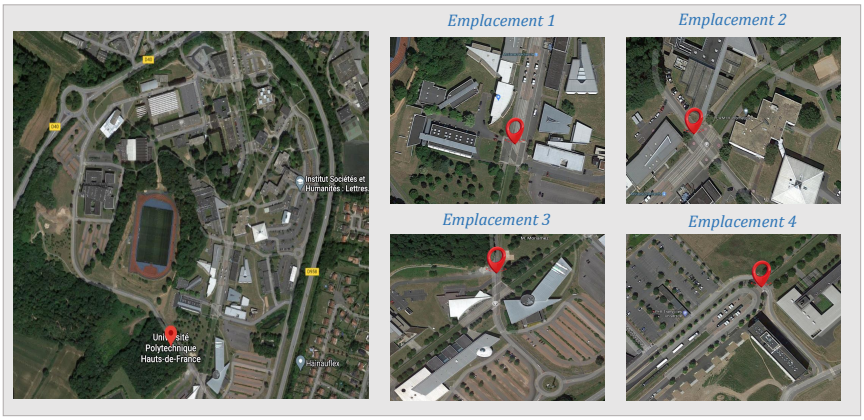


Figure 3.5: Recording emplacements at the University Polytechnic Hauts-de-France

3.3.7 *Dataset challenges*

With the intention of developing a complete dataset, we cover realistic conditions for environment perception (such as: cluttered environment, occlusions, lighting conditions,

etc.). Thus, we employed several sensors to obtain redundant information or complementary data that may compensate the challenges evoked by each sensor. Figure 3.6 highlights the most introduced challenges in our dataset.

- ☐ The object's types exhibit an immense variability since they vary in terms of appearance, movement and differ from the point of view of the class: pedestrian, vehicle, etc. When recording our data, we take into consideration this camera-radar challenge as we consider 4 categories of obstacles. Furthermore, our dataset was performed by several pedestrians and cyclists of different ages, looks, body sizes, etc. Moreover vehicles are varied: multiple cars, vans and trucks. We can see this differentiation through UWB radar signatures shown in figure 3.7 that correspond to each of the considered categories. Moreover, the considered objects can be static or dynamic.

- ☐ Distance is one of the fundamental challenges presented for autonomous driving either for camera, the two exploited radars or even the microphone. According to this, we consider two representations when capturing our dataset, depending on the range: near and far obstacles.

- ☐ A further challenge is presented : the cluttered environment since generally dense urban driving involves many traffic agents with a complex background. For UWB radar, multiple reflections can influence the quality of the signal in the presence of many objects. Concerning narrow-band radar, it generates many detections when various obstacles exist, thus a selection process is required to identify the relevant ones. So, we attempt to introduce several complex scenes during recording.

- ☐ Furthermore, we consider diverse lighting conditions as we record data throughout the day (morning, afternoon and sunset). We collect our dataset under sunny, foggy and snowy weather to increase the diversity and cover the possible real driving situations. In fact, the camera is highly sensitive to the last mentioned challenges whereas the radar is robust against them.

- ☐ Besides, the object detection task is extremely delicate to occlusions that occur between several classes which is frequently presented in diverse cluttered scenes. OLIMP includes severe occlusions situations combining the four classes as pedestrians that are often occluded by each other or by a cyclist, a vehicle or a tram, or the opposite.

Furthermore, the inter and intra class challenges are depicted in figure 3.8 by presenting the synchronized data acquired from the camera, the UWB radar and the microphone.

Figure 3.6: Challenges presented in our dataset: a) weather conditions, b) lighting variation, c) occlusions, and d) object types

3.3.8 *Statistics and dataset organisation*

OLIMP is organized in 6 subsets from C0 to C5. C0 contains background only, C1 includes either one, two or a group of pedestrians. C2 comprises cyclists, C3 and C4 include respectively vehicles and trams. The final subset C5 contains the different possible combinations of the aforementioned classes introduced in OLIMP dataset considering various scenarios. In fact, we only focus on the main moving road objects that can be presented in an urban traffic scene.

The dataset consists of **407 scenes**, and the number of scenarios in each subset vary as follows C0: 12 scenarios, C1 :144 scenarios, C2:31 scenarios, C3: 51 scenarios, C4: 18 scenarios and C5: 151 scenarios.

Our dataset was performed by 93 pedestrians, 14 cyclist and using 90 vehicles and 2 trams. Precisely, the dataset presents **47354 data for each sensor**. For the evaluation protocol,$\frac{2}{3}$ of the dataset is used for training, and $\frac{1}{3}$ for test.

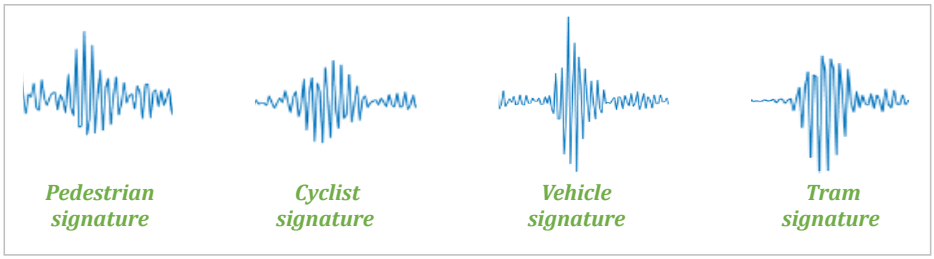


Figure 3.7: Object signatures extracted from UWB signals (near obstacles)

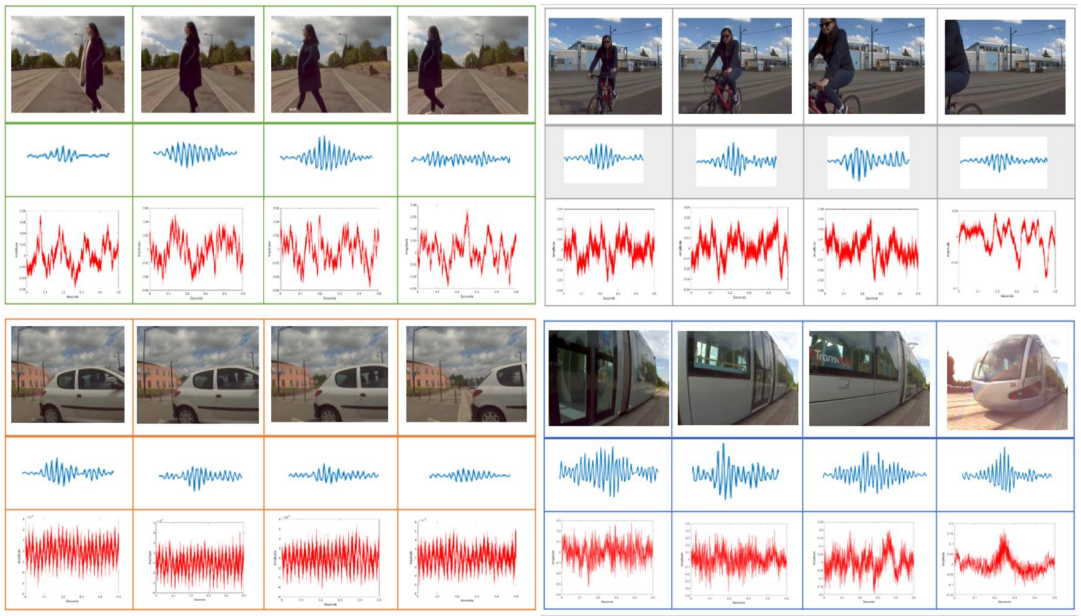


Figure 3.8: Inter and intra class challenges

3.4 FUSION FRAMEWORK

In order to propose an accurate multi-modal system for obstacle detection, in this section, we aim to evaluate each modality individually in order to propose, afterwards, a fusion-based system that takes advantage of each modality contribution.

3.4.1 *Image-based system*

The multiple obstacle detection task can be divided into two steps: the recognition that is insured via a probability estimation and the localization that defines the bounding boxes. Thus, deep learning techniques have been widely adopted in image-based object detection as it has been mentioned in chapter 2.

Among the known deep architectures used in the literature, we used the pretrained MobileNet-v2 [147] model on a subset of the ImageNet dataset for detecting objects on RGB images.

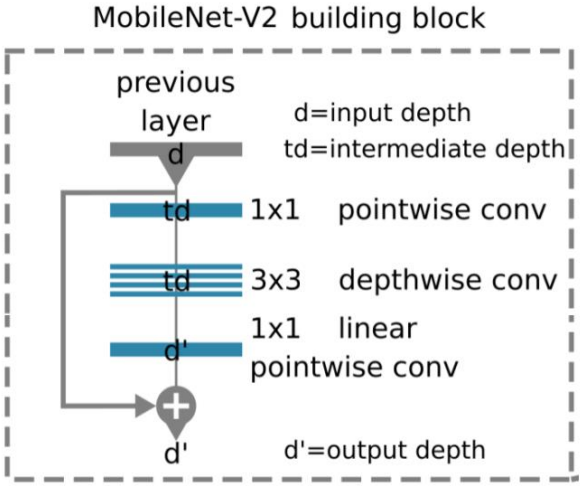


Figure 3.9: MobileNet-v2 building block

Inspired by the constraint of computations for mobile devices, the the light-weighted MobileNet networks was introduced in 2017 by presenting the MobileNet-v1 [128]. The next version is MobileNet-v2 which further improves the previous released version by employing a mobile inverted depth-wise convolution together with a residual connection. The MobileNetV2 is based mainly on depthwise separable convolutions and it contains two blocks. The first block is a residual block with a stride equal to 1 and the second block is a downsizing block with a stride equal to 2. Its architecture contains three convolution layers for both mentioned blocks: 1x1 convolution layer with RelU6 named pointwise conv, a depthwise convolution and a 1x1 pointwise convolution. The MobileNet-v2 building block is illustared in figure 3.9. The overall MobileNetV2 architecture contains 17 of these blocks. These blocks are followed by a regular convolution layer, an average

pooling layer and a fully connected classification layer.The network consists of 54 layers deep and uses 3.5 millions of parameters [147]. In fact, the presiding model was chosen due to its compromise between performance and execution time [128].

The results relating to the training of this network are presented in figure 3.10, and the metrics that are chosen to evaluate the performance are P, R and AP. The mAP reaches 60,5 %.

As shown in figure 3.10, MobileNet-v2 achieved a significantly higher results on the four categories in terms of precision. However, the image-based system provides high rates of recall for all the classes which explains that the system generates too many false negative samples.

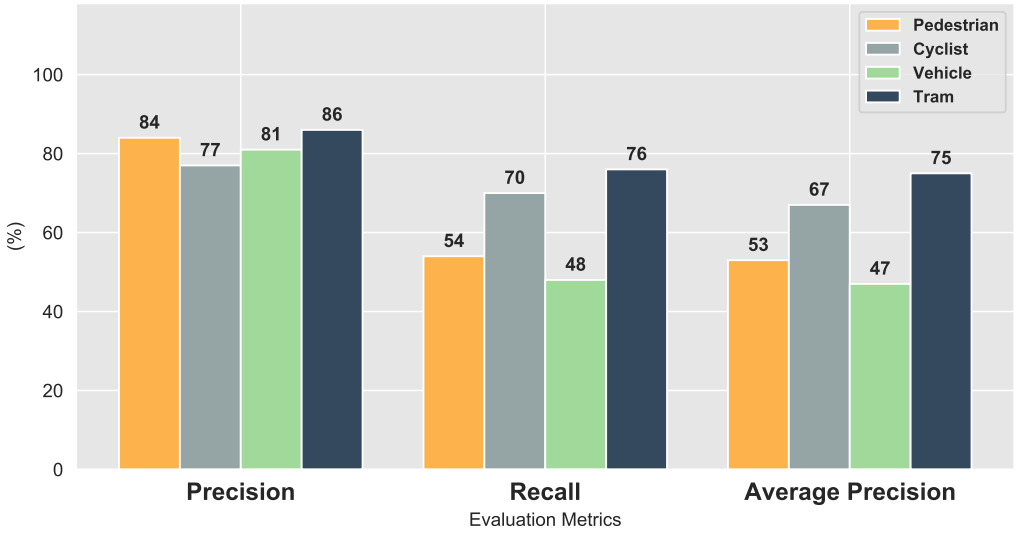


Figure 3.10: MobileNet results (%)

3.4.2 *Radar-based system*

To demonstrate the importance of using the UWB radar, we proposed a radar-based system to discriminate the four classes for short distances. First of all, we classified the whole signals using SVM in the intention of distinguish the classes, yet, the results were not promising as the signals present rich information with a significant leakage in the beginning. For this reason, we decide to exploit the narrow-band radar data to achieve better performance. Though, the proposed approach consists of selecting ROIs

in the signals acquired from the UWB radar in order to localize radar signatures that characterise the obstacle. Afterwards, these ROIs will be classified using SVM.
In fact, narrow-band radar generates a list of targets with their position and velocity. Thus, we injected the distances taken from narrow-band radar data to define the ROIs in UWB signals. In this state, we focus our attention to obstacles which are located less than 6 meters, while after various experiments the UWB radar is less efficient for a range that exceeds this margin.

We can observe that we obtain multiple ROIs when matching the narrow-band points with the signatures as the acquired radar are is too sparse. The detected distances are too close and may refer to a same object. Thus, in the aim to select the relevant distance and minimize the ROIs number, we proposed to exploit the velocity of each obstacle with the distance. This leads to a better localising of the signature. For that, two objects that are side by side and have the same velocity are considered as one target (represented in green color) as shown in figure 3.11.

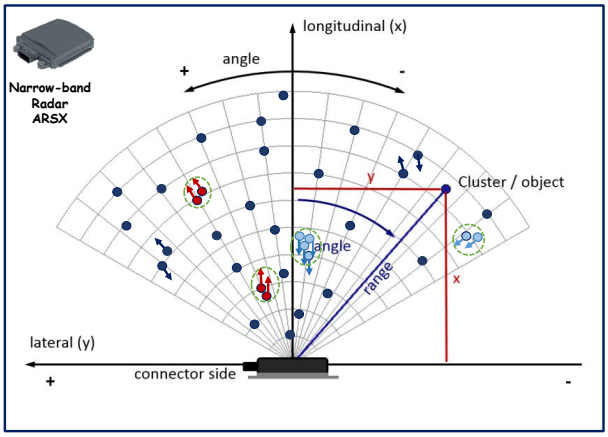


Figure 3.11: ERadar point cloud representation with optimisation using velocity.

In addition to this, we set an amplitude threshold to validate the ROIs. Figure 3.12 illustrates this process.
The selected ROIs are classified using an SVM classifier with an Radial Basis Function (RBF) kernel. The results of the UWB radar-based system are shown in figure 3.13.

According to our experiments and obtained results, we assume that the proposed radar-based system can better distinguish pedestrians and cyclists. Aside from the fact that the UWB radar provides a unique signature for each class, it is not able to classify tram and vehicle. Since the results in table 6 include the overall dataset testing, the accuracy

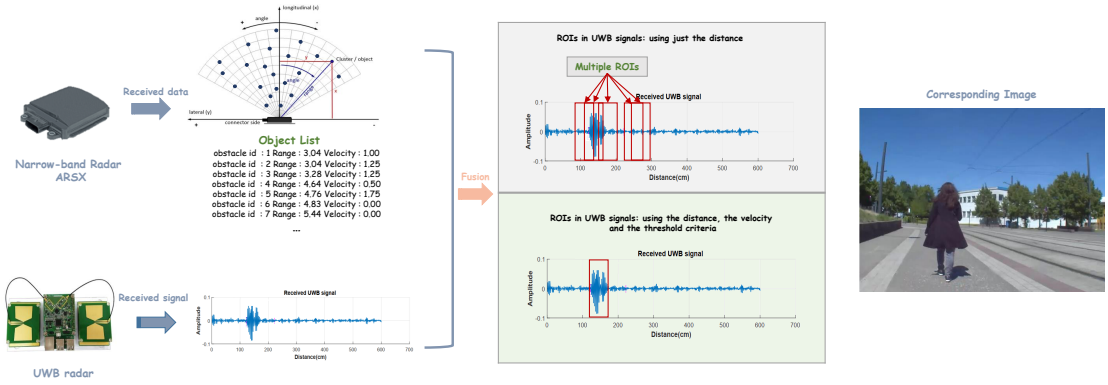


Figure 3.12: ROIs selection using UWB and narrow-band data.

results for those two classes are remarkably low. For experiments safety,the tram and the vehicle are generally located a far from the radar, in a range greater than 6 meters. Thus, reflections' magnitude from these two classes are low compared to reflections acquired from a cyclist or a pedestrian that are usually closer to the field of view of the radar. This explains the difference of accuracy between the two latter classes and the first classes.

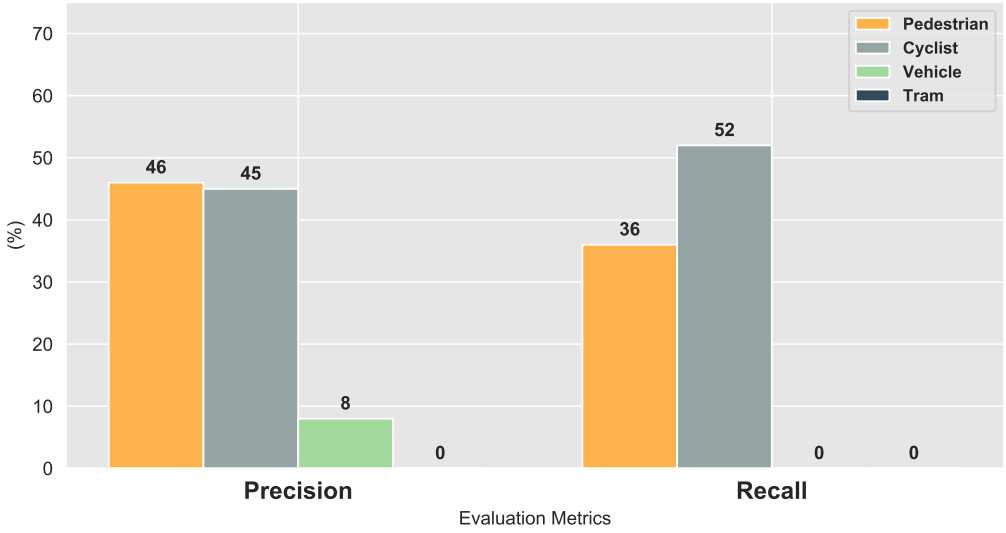


Figure 3.13: Radar-based system results

3.4.3 *Acoustic-based system*

According to the state of art, the MFCC (Mel-Frequency Cepstral Coefficients) are widly used in sound processing and analysis as it provides a better representation of the sound [151]. Hence, for acoustic data, we extracted temporal features and spectral features using MFCC based on several experiments. These features are concatenated and classified using SVM with RBF kernel.

As shown in the results presented in figure 3.14, using acoustic data leads to better performance for the two categories tram and vehicle. This is due to the relevant sound generated by these two classes. In other words, a walking pedestrian sound is narrow compared to the tram sound that presents more information. For this reason, precision and recall rates related to the tram and the vehicle classes are higher than the two others.

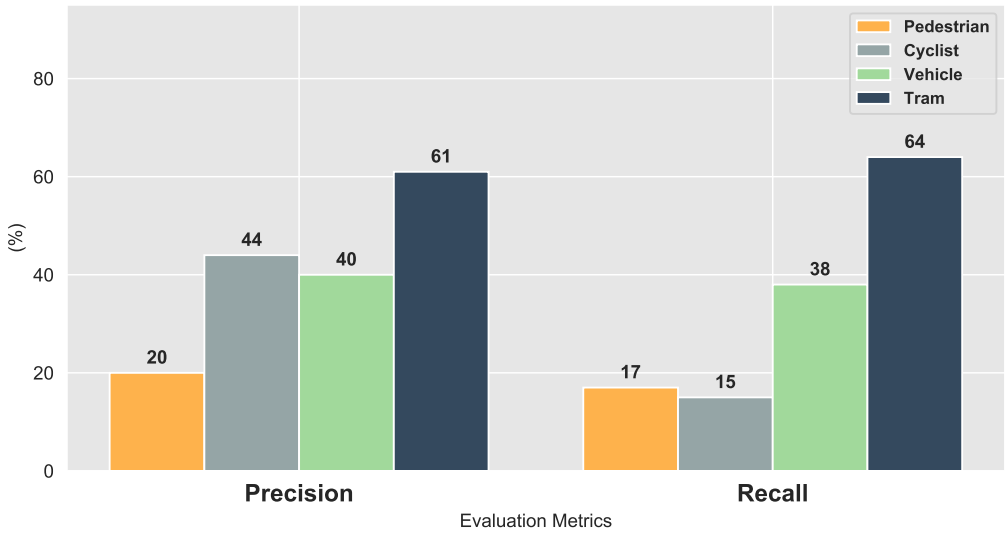


Figure 3.14: Acoustic-based system results (%)

3.4.4 *Multi-modality fusion system*

To prove the significance of our dataset and the importance of multi-modality aspect, we take advantage of the different sensors by proposing a fusion framework system. This framework is built in the lights of the results obtained from the aforementioned

systems. In fact, we identify the effectiveness of each sensor individually and its ability to differentiate one class of another according to the results presented in Section 3.4.1-3.4.3. The architecture of the proposed fusion framework is represented in figure 3.15.

The first step of the framework consists of extracting the labels from the outputs of the MobileNet. If the extracted label is a car or a tram, we use the acoustic-based system to verify the attributed label, and, all the labels are updated accordingly. Subsequently, if the CNN-extracted label is neither a tram neither a car, the distance of the object will be calculated. The process of calculating the object distance is explained in the following. Thus, if it is a far obstacle we will keep the same labels of the CNN model. Nonetheless, if it is a near obstacle then it will be either a pedestrian or a cyclist. Along with, we will adopt the radar-based system to confirm the attributed label, since it can particularly discriminate the aforementioned categories in a range less than 6 meters. Thus, the results related to the fusion framework are illustrated in figure 3.17.

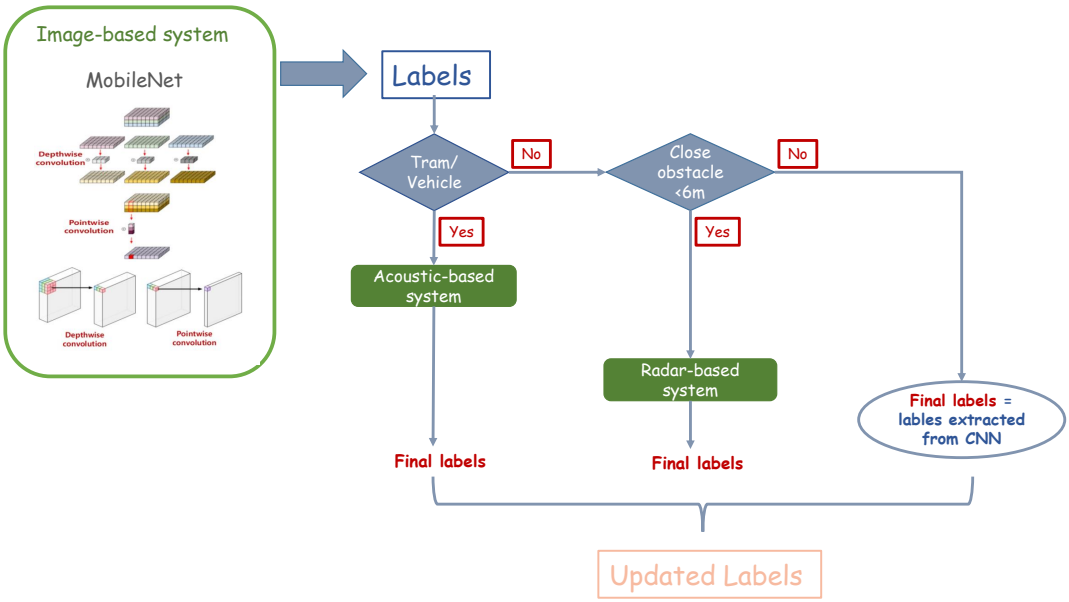


Figure 3.15: Proposed fusion framework architecture

☐ Distance calculation using camera images : In the aim to find a way to select suitable object's distance, We adopt to use the bounding boxes obtained after the testing process using MobileNet-v2 network.
In fact, to perceive the depth information from images, stereo-vision techniques are generally used. However, they require an intensive computation and in our case we deal with a monocular camera. For this reason, we propose to exploit an *area-based* approach to determine the object distance from the available bounding

boxes. In other words, we look to find a relation between the object's area in the image and its real distance.
Therefore, the area of the bounding boxes is estimated at several locations for each obstacle type and are stored in a training dataset. By using a curve fitting and optimization techniques, the data give a non-linear relationship between the area of the bounding boxes of the images and the real distance of the objects.

The relationship between the area of the bounding box and the object's distance can be modulated via Equation 3.1.

$$Object_{distance} = \alpha(Area)^b + c \tag{3.1}$$

where the relative parameters are : α = 741.3 , b = -0.507 and c = -0.3258. This process is illustrated in figure 3.16.

It should be pointed that the obtained relationship between the area and the distance using images is only valid for our employed camera.

Figure 3.16: Estimation of object distance using area of detected bounding boxes.

3.4.5 *Discussion*

On the one hand, we conducted various experiments using mono-modality and multi-modalities to validate our dataset and to open perspectives the way for future research. On the other hand, these experiments show the significant impact of exploring multi-modality and data fusion for an ITS to improve the obstacle detection task.

In fact, the fusion levels exploited in our work are the following: low level, intermediate level and late level. We can recognize the *low level fusion* when projecting narrow band

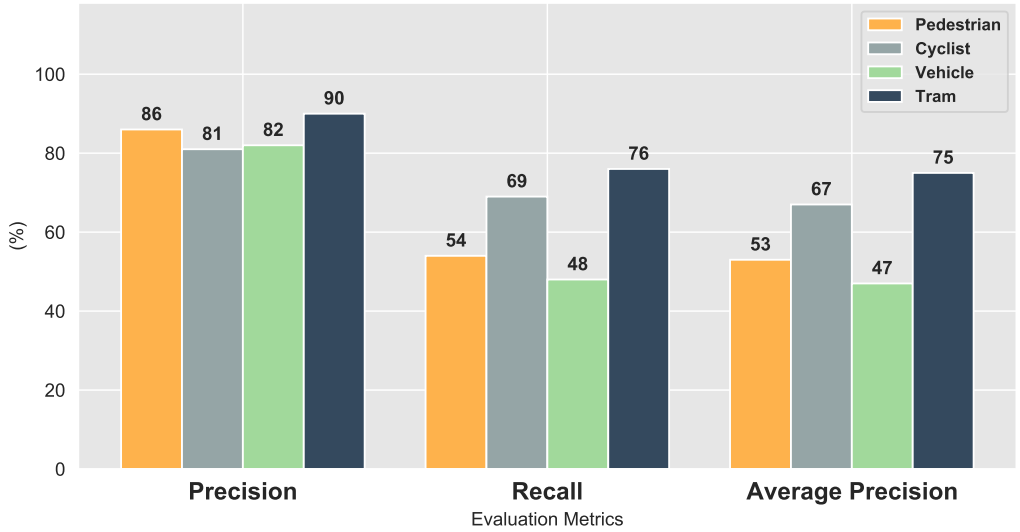


Figure 3.17: Fusion framework results (%)

data into UWB signals to define ROIs. The *intermediate fusion* consists of concatenating temporal and spectral features for acoustic data. And the *late level* is exploited in decisions fusion to obtain the final decisions of the total framework.
From analysing the fusion results presented in figure 3.17, we notice that the performance has been clearly improved in terms of precision. Some detection results that include bounding boxes and scores are presented in figure 3.18 using just image-based system and with using fusion framework. The enhancement brought along with the acoustic system has a higher importance compared with the contribution of the radar-based system. This is mainly because of the range and power limitations of the UWB radar, and also owing to the sparse and noisy data provided by the narrow-band radar. Despite this fact, the UWB radar provides a unique signature for each type of object with a low price compared to the new sophisticated radars for short range applications. For this reason, this radar is more explored in the next chapter. Thus, we will focus especially on improving this single-sensor performance because it carries rich information, and due to its important short-range settings aspect. For the acoustic system, the distance between the obstacle and the sensor presents an important challenge. Moreover, obstacles like pedestrians and cyclists have low magnitude acoustic signals and could not easily detected through acoustic based systems. In addition, The considered environments in OLIMP are challenging and present various confusing categories such as metal infrastructure, trafic signs, glass-surface buildings, etc.

Finally, The obtained results for object detection are promising and show the importance of exploiting multimodality for vehicle environment perception. To the best of our knowledge this is the first dataset that has exploited UWB technology and acoustic data, this shows the originality of our work. For this reason, we encourage research on proposing new fusion networks that use either two modality or more to enhance the vehicle environment perception.

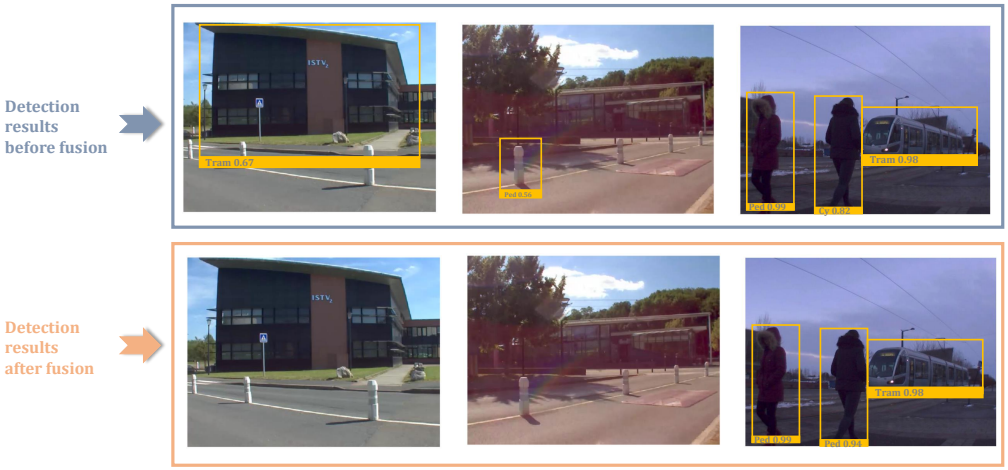


Figure 3.18: Detection results before and after fusion

3.5 CONCLUSION

In this chapter, we have presented a state-of-the-art of available datasets that were introduced for environment perception. By analysing the characteristics of each dataset, we have concluded that there is no dataset that employs UWB radar and acoustic data. Following, we represented our multi-modal dataset by detailing its hardware, its scenarios, employed sensors, etc. To validate our dataset, we conducted various experiments using mono-modality and multi-modalities, then, we proposed a fusion framework that enhances the detection performance.

In the next chapter, we will focus on detecting objects using just the UWB radar. Thus, two detectors will be proposed one based on entropy and the other a deep-based detector.

4

MULTIPLE OBJECT DETECTORS USING UWB SIGNALS

Contents

4.1 INTRODUCTION

The development of high performance obstacle detection systems is a safety critical task for intelligent and autonomous systems. Therefore, various sensors are employed to efficiently detect and recognize objects. Due to its robustness to weather conditions, radar is a promising sensor for environment perception systems. The main radar technologies that have been exploited in this domain are wide-band radars for short range applications and narrow-band radars for long range settings. In this chapter, we focus on the UWB sensor as it carries rich information.

In section 4.2, the UWB radar specifications are depicted. An overview of existing hand-crafted and deep learning-based methods using UWB signals are presented in section 4.3. In section 4.4, a novel entropy-based method is proposed for UWB radar-based multi-target detection and we discuss our results and their potential impact. Following, we propose the first framework that exploits LSTM with UWB signals for multi obstacle detection in an outdoor complex environment. Afterwards, we exhibit a comparison of our results with the state of the-art-techniques in section 4.5.2. We conclude the chapter in section 4.7.

4.2 UWB RADAR SPECIFICATIONS

In 2002, the US Federal Communication Commission (FCC) allows the unlicensed of UWB operations and commercial use of UWB-based devices [27]. Firstly, intended for military purpose applications, the UWB radar has been exploited in various applications. In fact, three main applications are defined *(i)* communications and measurement systems [40], *(ii)* imaging systems as through-wall imaging systems [154], and medical systems [1], *(iii)* vehicular radar systems [92]. The FCC permits a frequency range of UWB of 3.1 to 10.6 GHz in order to avoid the interference along the existing communication systems [27]. In addition, it defines that the UWB signal should has a fractional bandwidth more than 500 MHz or characterized by 20 % of the center frequency.

The UWB radar transmits a very short electromagnetic pulses with low energy in the order of sub-nanoseconds. These short pulses provide a huge interest in short range radar applications [144]. Various types of waveforms are used to generate the UWB pulse as the Gaussian and its derivatives. Besides, the Gaussian monocycle is the commonly type that is utilized as UWB impulses.

Basically, UWB radars can be categorized into two groups: pulse-based radars (e.g. time domain) that are typically referred to impulse-radio UWB (IR-UWB) and frequency modulations-based radars refereed as CW UWB radars (e.g. frequency domain) such as the frequency modulated continuous wave (FMCW). A comparison of the two technologies is provided in Table 4.1. In our work, we are interested in IR-UWB radar.

Table 4.1: Comparison between CW-radars and IR-UB radars [99]

UWB radar Technology	**CW UWB radars**	**IR-UWB radars**
Time stablility	Stable	Stable
Availability	Require highly linear signal generator	Available and a widely used technology
Cost	High cost	Inexpensive technology
Data acquisition Time	Slow data acquisition (ms)	Moderately fast data acquisition (μs)
Interference Immunity	Relatively immune towards narrow-band interfernce	Relatively immune towards narrow-band interfernce
Dynamic range	They possess very good dynamic range	Might have a problem maintaining linear dynamic range
Power	Good power budget	Low average transmitted power
Range gating	No possiblity of range gating	Allow range gating

In fact, the UWB radar has interesting characteristics for ITS applications as it has a better resolution than existing narrow-band radar devices. It consumes low power, it has a simple implementation and has a high data rate commutation. The UWB radar enables the penetration in dielectric materials. Therefore, the major property of the UWB radar consists of the distortion of the initial pulses. This signal deformation is impacted by the obstacle properties and thereby represents the object signature. This signature contains information that goes beyond the distance and the velocity; it is shaped by the object material, its shape and size [143]. The UWB is able to detect stationary and moving obstacles on the vehicle's surroundings (on and nearby the road).
The target's distance R is measured based on the delay between the emission and reception (τ), c is the speed of light. This relation is presented by Equation 4.1. The target's range calculation is illustrated in figure 4.1.

$$R = \frac{\tau c}{2} \tag{4.1}$$

These aforementioned characteristics show that the use of such radars is promising in detecting and recognizing objects, especially in short range applications, contrary to

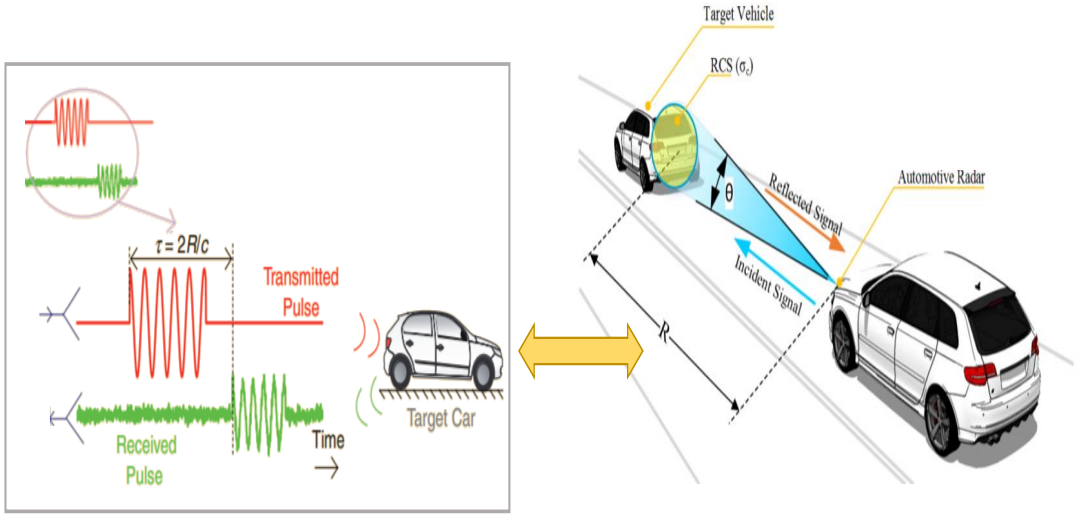


Figure 4.1: Target range calculation using automotive radar.

narrow-band radars that detect range targets with low accuracy and generate sparse data that contain numerous false alarms [99].

While the UWB radar offers rich information that is complementary to other sensors, its deployment presents serious challenges. One of the main challenges is differentiating targets from noise, which can practically be formulated as a segmentation problem.

4.2.1 *UMAIN radar*

The radar considered to record the OLIMP dataset, is an UWB radar developed by the UMAIN Inc company [163] entitled HST-D3. It has an efficient range of 6 meters and a frequency range in [3GHZ, 4GHZ] with a bandwidth of 0.45-1Ghz.
The HST-D3 radar is a combination of HST-S1 Pi module radar and a Raspberry Pi 3 for the acquisition. It is a high-resolution radar transmitting and receiving UWB impulse on a single chip. Moreover, it has a Baudrate of 921600 and the UWB radar signal comprises 660 samples per frame. Total frame time is renewed every 22.5 ms and the interval per value in a single frame is about 2.0303 cm.
This radar provides implemented algorithms as respiration detection, human and animal detection. In addition, the user can develop and implement its own algorithms.

The UMAIN radar contains an UWB monopole and UWB directional antennas. We exploit the available directional antennas since they guarantee better target echo-to-clutter and noise ratio. The radar is presented in figure 4.2 and table 4.2 exposes the radar specifications.

In fact, as we mentioned in section 4.2 UWB signals have excellent multipath immunity, and less susceptibility to interference acquired from other radios, due to its wide bandwidth property [59]. However, the limitation of the UMAIN frequency range is that interference can appear due to WiMax(Worldwide Interoperability for Microwave Access) technology-based applications. Concerning the atmospheric attenuation, it is negligible for short ranges fields using the S-band [36].

Table 4.2: Umain radar specifications

Parameter	**Value and comments**
Frequency range	3∼4 Ghz
Bandwidth	0.45∼1Ghz
Output Power	Typ. -25dBm
Antenna Specification	UWB Directional Antenna : Gain = Avg.7 dBi Antenna angle (@-3dB) = 56.0° (X-Z plane) 77.5° (Y-Z plane) Size= 76mm x 58.5mm x 17mm
Number of samples	660 samples per frame
Sampling frequency	7,69Ghz

4.3 RELATED WORK

4.3.1 *Hand crafted-based detectors using UWB radar*

Most of the studies conducted on UWB-based obstacle detection systems have taken advantage of the well-known of three mainly approaches that have been proposed in the literature: correlation-based method as the matched filter [91], higher order statistics (HOS) [104] and constant false alarm rate (CFAR)[53]. These techniques are known as hand-crafted based detectors.

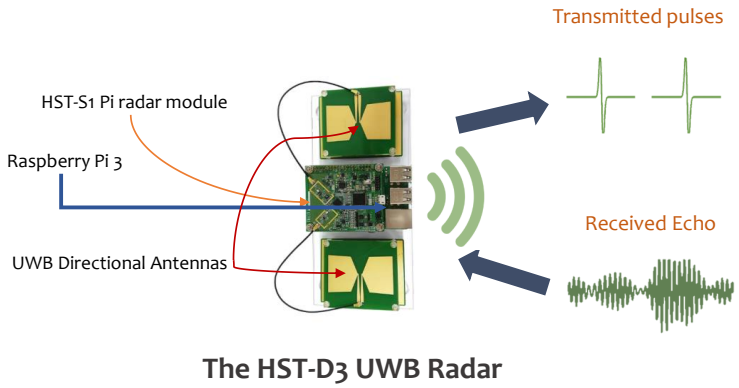


Figure 4.2: Used HST-D3 UWB radar hardware specifications.

HOS technique relies on higher-order moment spectra to analyse random process characteristics. This technique is commonly used to suppress Gaussian noise [104].

CFAR techniques [53] [116] detect objects using an adaptive threshold that is defined based on the background noise and local information. Several algorithms have been extended from CFAR including the cell averging CFAR (CA-CFAR) [28], the order static CFAR (OS-CFAR) [12], the Smallest Of CFAR (SO-CFAR) [21], etc.

In [146] the correlation is used to detect either a car, a metal plate, a motorway barrier or a pedestrian. A recent adaptive clutter suppression algorithm is proposed in [187] based on CFAR technique, for human detection and positioning. An hybrid method was introduced in [141], where a radar target detector based on the combination of HOS and CA-CFAR techniques was deployed. Nevertheless, experiments were performed under controlled scenarios. A new thresholding method based on CFAR technique for UWB-based detection application is proposed in [129]. The approach takes into consideration false alarm and miss-detection criteria. New parameter entitled constant miss-detection rate (CMDR) is defined. The final threshold is calculated by adding the CFAR and CMDR rates. Experiments show good performances, however, they are just carried out in indoor environment using self-recorded dataset. A recent work in [82] proposes UWB radar-based system to detect metal lane. The energy of the echo is calculated and if it exceeds a fixed threshold, the lane is detected. The use of this method requires a specific infrastructure. Moreover, energy values are dissimilar for different object types compared to metal obstacles energy.

In fact, almost of the aforementioned related work are a threshold-based methods. Otherwise, these techniques depend essentially on the amplitude of the object's signature. Moreover, it should be pointed that the considered environment and scenarios are controlled using restrained datasets. The aforementioned algorithms have been exploited ever since the development of the conventional radar.

4.3.2 *Deep learning-based detectors using UWB radar*

To the best of our knowledge, UWB-based systems that rely on deep learning techniques are employed only for indoor applications as: activities recognition [142], indoor people localization [123], recognizing movements during sleep [122], etc.

Deep learning methods have been used with 2D UWB data, thus, this dilemma is considered as an image processing-based challenge. In [24], an SFCW-UWB radar is used for fall detection which generates time-frequency spectrum as UWB data. Hence, the employed deep learning architecture is the fine-tuned Alexnet model. The authors in [85] convert the time-series UWB data to a time-frequency representation by Stockwell transform. Afterwards, the reshaped images serve as inputs of the CNN. The exploited CNN model is LeNet. In [1], UWB data are stored as 2D matrix including the slow-time and the fast-time properties. Subsequently, it is converted to a grayscale image, later to an RGB one. GoogLeNet was adopted as deep learning architecture for hand gesture recognition. In [122], weighted range-time-frequency images of UWB radar are utilized as input for CNN to classify human sleep postures.

In point of fact, the 2D radar imaging based systems are no longer considered a signal processing dilemma as much as an image processing challenge.

Otherwise, 1D UWB signals are likewise employed with deep learning for indoor applications.
To enhance transportation safety, an UWB radar is installed in the rear view mirror to estimate the number and the location of the in-vehicle people. Multi-layer perception is employed where the time-sampled radar signal data are the input of the network. To define the suitable parameters, the number of hidden layers are adjusted. Compared to machine learning techniques, the proposed network achieves better results [92]. For activities recognition, a CNN-LSTM network using three UWB radars is implemented in [102]. Features are extracted using CNN architecture that includes: two 1D convolution layers where 64(1x3) filters are used, Relu activation function, a 1D max pooling layer. Subsequently, the output is flattened into 1D vector to feed the LSTM network that

contains 2 LSTM layers. Promising results are obtained. Jiang et al. [77] trained a CNN-LSTM to classify Line-Of-Sight (LOS) and None-Line-Of-Sight (NLOS) signals in the context of indoor positioning applications. The UWB channel impulse response is used as input of the CNN that deploys two conventional layers. Afterwards, the CNN outputs are linked to the LSTM network. Bi-direction LSTM and a stacked LSTM are used. The achieved accuracy is equal to 81%, but, the training dataset was limited.

The aforementioned applications consider only indoor environments, using either a 2D radar imaging or a 1D UWB data. Nevertheless, ITS environment includes complex driving situations with various type of targets. Moreover, in our case the choice of adopting 1D radar signal rather than using a 2D radar data representation is justified by the fact that we deal with an ITS application where response time is a crucial criteria.

4.4 ENTROPY-BASED DETECTOR

4.4.1 *Background*

While UWB radar has been widely used for obstacle detection, most of the reviewed works are evaluated on restrained datasets that have been recorded under controlled indoor environment. Furthermore, these techniques rely on the signature amplitude, which depends on the distance between the radar and the object location. This represents a fundamental challenge for UWB signals segmentation using amplitude-based techniques. This challenge is even more critical for human obstacles because of their low reflection compared to metallic obstacles.

Thus, the fundamental challenge of the detector is to distinguish a target from a received additive noise. To tackle this challenge, we suggest to use the signal entropy as an indicator of the existence of useful portions of the signal that can be differentially extracted out of the channel noise. So, in our case, we aim at segmenting 1D UWB signal used for an outdoor setting with a complex environment.

Recently, entropy-based information using UWB technology has been exploited in various fields of research as telecommunication[186], medical[15], etc. In [101], mutual information (MI)-based methods are applied to detect targets through foliage based on the calculation of entropy and conditional entropy. The defined threshold is log2(level of quantization of the signal). In fact, if the MI of the received echo exceeds the threshold, the target is detected. The proposed approach in [182] adopts the permutation entropy

(PE) for detecting human vital signs. As the PE detects dynamics changes in time series signals, it is employed to determine the range between the radar and the human target. For people-localization-based on UWB technology, the Shannon's entropy is used to accurately estimate the time of arrival of the first path in indoor environments. It is detected by identifying a great decrease in the entropy curve, and that exceeds a threshold value which has been defined via numerical simulation[186].

While the entropy is a measure of the signal's complexity, different types are employed in the literature as: PE, conditional entropy, Shannon entropy, etc. In fact, every type of entropy is suitable for a specific application and data. For our application, the entropy is applied as a segmentation tool for UWB signals in outdoor environment.

From an information theory perspective and as defined by Shannon [153], the Shannon's entropy is a theoretical metric of information. It is attributed to an information source or a given signal, and models the amount of information contained in this source or vehiculated by the signal. The Shannon's entropy may be used globally or locally: taking into account the whole data or a subset of data [16]. Therefore, to localize objects within an UWB radar signal, the use of the Shannon's entropy should be helpful in distinguishing useful ROIs from noise. This is based on the assumption that compared to noise, reflections from road obstacles present richer information. An illustration of this assumption is given in figure 4.3. The figure represents a received signals from two different objects (a cyclist and a vehicle) and its corresponding entropy curves. The bounded region (in purple) within the received signal represents the correct ROI which is depicted by the highest entropy value.

Thus, our hypothesis is that the signal corresponding to an obstacle has potential information that is different from a random Gaussian noise. Thereby, the entropy distribution within each part of the signal should be different. Therefore, we suggest to exploit a differential entropy analysis of the received echo to localize ROIs within the signal.

In the following, we first provide a theoretical basis that backs the use of entropy for segmentation. Then, we detail our segmentation method.

4.4.2 *Theoretical basis*

In this work, our hypothesis is to use the signal's entropy as an indicator of the existence of useful parts of the signal (i.e., parts corresponding to obstacles in the radar's neighborhood) that are distinguishable from noise. Accordingly, we attempt to show that

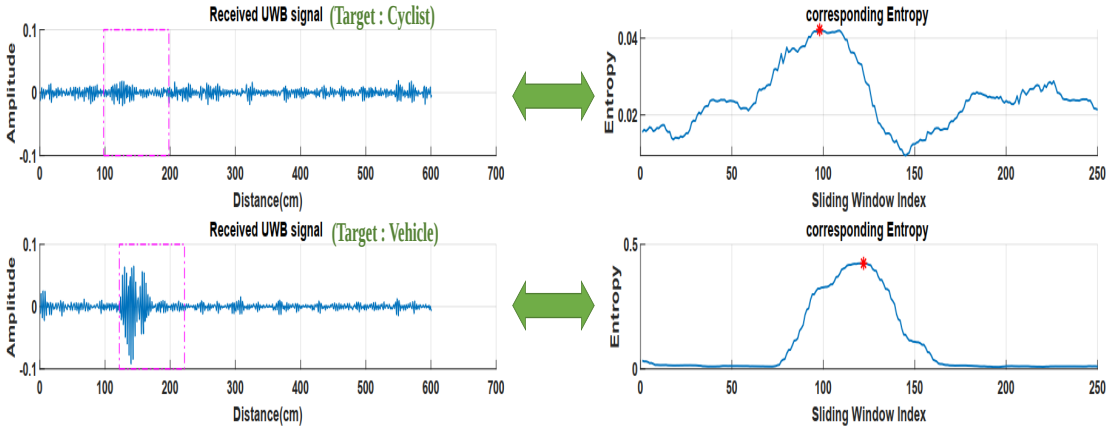


Figure 4.3: Illustration of the entropy variation of two different received UWB signals.

statistically, a useful signal should contain higher entropy than a noise signal. Notice that this observation will practically depends on different real-life parameters. Hence, in this study we make the following assumptions:

- ☐ Since UWB radar is mainly used for urban traffic situations because of its short range, we assume a propagation model of an UWB radar that follows Rayleigh distribution, as shown in [108].
- ☐ We assume an Additive White Gaussian Noise (AWGN) in the propagation channel of the UWB radar.

The received UWB signal $r(t)$ can be modeled following Equation 4.12:

$$r(t) = s(t) + n(t) \tag{4.2}$$

where $s(t)$ is the received echo and $n(t)$ is the noise of the transmission channel of the UWB radar. In fact, the process of emitting and receiving an impulse by the UWB radar is labeled as a radar scan, and the received echo of the received signal of the i^{th} radar scan $r_i(t)$ can be modeled by Equation 4.3.

$$r_i(t) = \sum_{k=1}^{M_i} a_{ik} x(t - t_{ik}) + n(t). \tag{4.3}$$

where: $x(t)$ is the transmitted pulse that is received as M_i reflected signals, a_{ik} is the amplitude, t is the reflected back time of the pulse signal after being transmitted from the radar, t_{ik} is the delay of the k^{th} received signal in the i^{th} radar scan, $n(t)$ an additive Gaussian noise of the transmission channel.

Therefore, r_k can be observed with corresponding probabilities p_k $\forall k \in [1, L]$ and the entropy H is expressed by Equation 4.4.

$$H = -\sum_{k=1}^{L} p_k \log_2(p_k) \tag{4.4}$$

The Rayleigh and the normal distributions have been proven log-concave in [6]. Hence, based on the work on [13], the entropy could be approximated as: $H(X) \approx \log(\sigma)$, where X is a random variable following a log-concave probability distribution and σ is its standard deviation.

Based on a statistical analysis of the distribution of both noise and useful signals (i.e. those in the presence of an obstacle) from a real-world dataset (OLIMP) [106].
We found that $\sigma_{noise} = 0.0056$ and $\sigma_{obstacle} = 0.0416$. Hence, $log(\sigma_{noise}) < log(\sigma_{obstacle})$, as shown in figure 4.4.

Therefore, we can conclude that for a given UWB radar signal propagated in a Rayleigh fading channel with AWGN, we statistically have: $H(N) < H(U)$, where N and U are random variables that represent the noise signal and the useful signal, respectively and H is the Shannon entropy.

In the following subsection, we detail our proposed segmentation approach.

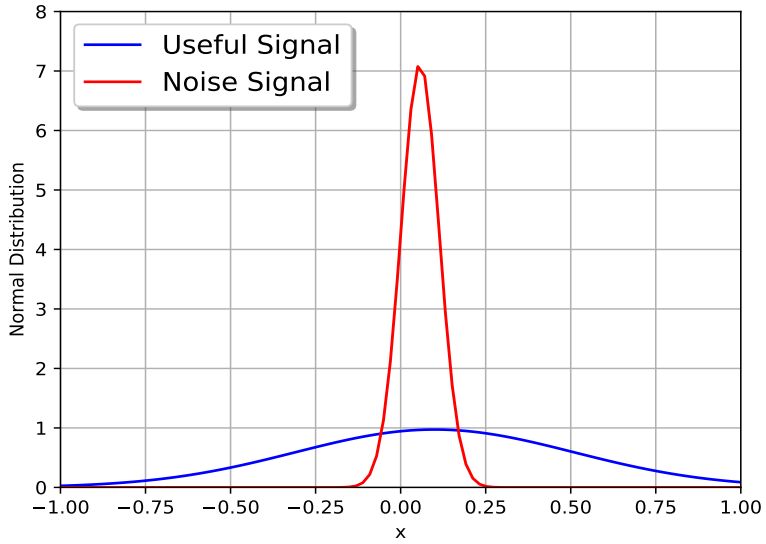


Figure 4.4: Comparative distribution of noise signals, and signals in presence of an obstacle.

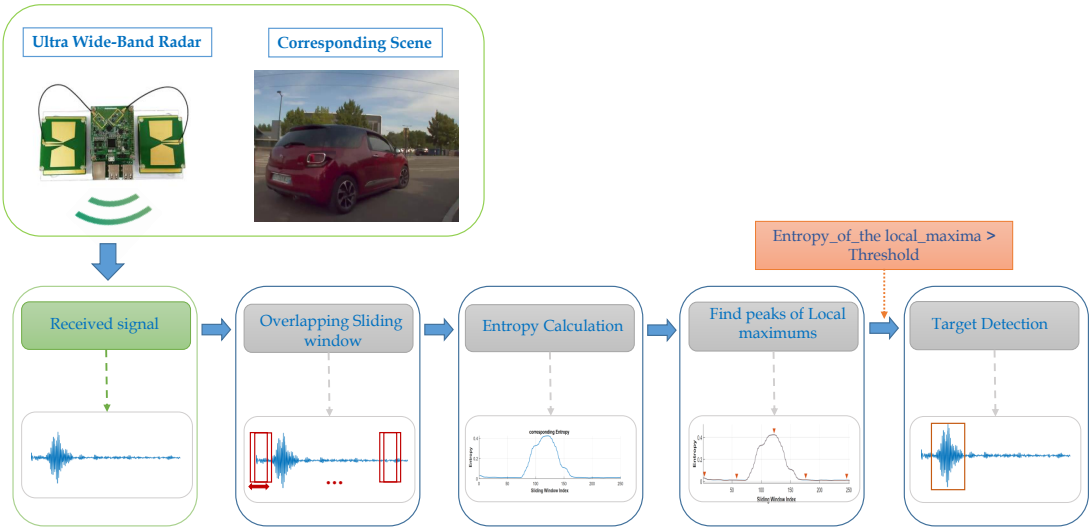


Figure 4.5: Entropy-based detector architecture

4.4.3 *Entropy-based segmentation*

This subsection explains the proposed method by which ROIs are detected within the signal. Algorithm 1 details the ROI identification process and figure 4.5 presents the architecture of the proposed approach.

The received echo is partitioned into fixed-size overlapping sliding windows. The sliding window's size (W_{slide}) is empirically defined based on the objects' signature length. As the overlapping sliding windows should not miss the signal's valuable parts, we slide the window by one sample at a time. Subsequently, the Shannon's entropy is calculated for each window. The variation of the entropy values is then obtained by sliding the window through the whole signal. Since the entropy's value increases relatively with the rise of the signal complexity (for example: presence of different objects or noise), we localize the local maximums presented in the entropy curve in a range of W_{slide} length.

Local maximums within the same W_{slide} are ignored to prevent false detections that can be recurrent for the same object. The obtained maximums represent potential candidates for ROIs. Some of these candidates may be false positives generated by the noise entropy.

Therefore, to detect only relevant entropy peaks, and by consequence limit false positives, a selection process is designed based on the study in Subsection 4.4.2. In fact, an empirical threshold is defined to withdraw noise-generated entropy peaks. Therewith,

if a local maximum of the entropy is greater than the fixed threshold, a ROI is detected. The candidate is considered as noise otherwise. The threshold definition methodology is detailed in Section 4.4.5.

Algorithmus 1 : Entropy-based segmentation technique

Data : Received signal $r = [r_t]$, Sliding window length: $Wslide$; Threshold: Thr;
Output : ROI

for $t \leftarrow 0$ **to** $length(r) - Wslide$ **do**
    $Sw = r(t+1, t+Wslide)$;
    `// Calculate the Shannon's entropy`
    $En_t =$ `Entropy`(Sw)
end for
`// (peaks, indexes) of local max.`
$[LmaxEn, IndxLmaxEn] =$ `Findpeaks`(En_t);
`// Compare LmaxEn with the threshold`
$ROI = \emptyset$
$\forall i \in IndxLmaxEn$
if $LmaxEn_i > Thr$ **then**
    `// Define the ROIs`
    $ROI = Append(ROI, [i : i + Wslide])$;
end if
return (ROI)

4.4.4 *Experimental setup*

OLIMP dataset contains four object classes: pedestrian, cyclist, vehicle and tramway, and it includes various urban driving scenarios. In the following experiments, we compare our approach to HOS technique, CFAR technique and the work in [141]. A brief description of these techniques is provided in the following.

(a) ***HOS***

The HOS algorithm relies on the higher order moment spectra in the intention of interpreting and analyzing the characteristics of a random signal.
One of the main advantages of employing this technique is its ability to reduce the Gaussian noise and the secondary lobes. Moreover, it characterizes and detects

the non linearity in the data [145]. Thus, Hos is used to detect various obstacles by applying a simple threshold. In this work, the 4^{th} order cumulant based on Tuganit4 algorithm is implemented for HOS method.
The Tugnait4 Algorithm was introduced by Jitendra K.Tugnait [162] in 1989 and it is expressed by Equation 4.5.

$$J_4(i_0) = \frac{\mathbf{cum_4}(c(i-i_0), c(i-i_0), r(i), r(i))}{\sqrt{|\mathbf{cum_4}(c(i), c(i))|\,|\mathbf{cum_4}(r(i), r(i))|}} \tag{4.5}$$

where: c is the reference signal, r is the received signal, i_0 is the decision time index and cum_4 is expressed by Equation 4.6.

$$\begin{aligned} cum_4(c(i-i_0), c(i-i_0), r(i), r(i)) = \frac{1}{N}\sum_{i=1}^{N-1} c^2(i-i_0)r^2(i)) \\ -2\left[\frac{1}{N}\sum_{i=1}^{N-1} c(i-i_0)r(i))\right]^2 \\ -\left[\frac{1}{N}\sum_{i=1}^{N-1} c^2(i-i_0)\right]\left[\frac{1}{N}\sum_{i=1}^{N-1} r^2(i)\right] \end{aligned} \tag{4.6}$$

(b) ***CA-CFAR***

For CFAR, the automatic threshold CA-CFAR detector is considered. The basic principle of this technique is shown in figure 4.6. In fact, the CA-CFAR is based on local information to adaptively define a threshold to detect the targets [52] [136]. The parts on the two sides of the Cell Under Test (CUT) are named guard cells, these cells do not participate when estimating the ground clutter power in order to avoid missing detection. The detection threshold T is expressed in Equation 4.7 [113]. The CA-CFAR detector decides whether there is an object or not by comparing the power of the CUT to the threshold.

$$T = \gamma Z = \frac{1}{2n}\gamma(P+Q) = \frac{1}{2n}\gamma\left(\sum_i^n p_i + \sum_j^n q_j\right) \tag{4.7}$$

where: p_i and q_i, $(i, j = 1, ..., n)$ are the samples of the reference cells on both positions of the CUT, P and Q are the power summation related to the front and

the back edge reference cells, Z is the average of all the reference cells and γ is the threshold coefficient that depends on the desired probability of false alarm rate P_{fa}.

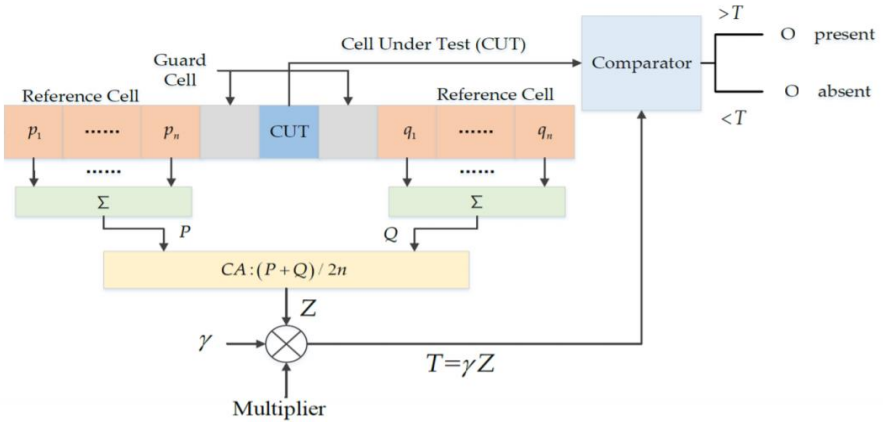


Figure 4.6: CA-CFAR algorithm architecture [113]

(c) ***The work in [141]***

The work in [141] is a combination of the HOS technique and the automatic CA-CFAR detector. In fact HOS is employed first to suppress the noise. Afterwards, the CA-CFAR is applied on the received signal following the elimination of the noise.

In our work, all the experimental results are evaluated in terms of P, R and F1-score that balances P and R, these metrics were presented in chapter 2 in section 2.4.

Moreover, to evaluate the performance we used the Multiple Object Detection Precision **(MODP)** metric. This measure evaluates the positions' precision of the accurately detected objects. The overlap information between the system's detection and the ground truth is used to calculate a Mapped Overlap Ratio (MOR) for each frame as defined in Equation 4.8.

$$MOR = \sum_{i=1}^{Nm} \frac{Gi \bigcap Di}{Gi \bigcup Di} \tag{4.8}$$

Where Gi is the bounding box of the i^{th} ground truth and Di is the bounding box of the i^{th} detection and Nm is the number of the mapped outputs. For a single frame, MODP is obtained by normalizing MOR as shown in Equation 4.9.

$$MODP(t) = \frac{MOR}{Nm} \tag{4.9}$$

Hence for a multiple frames, the score MODP is determined by Equation 4.10.

$$MODP = \frac{\sum_{i=1}^{Nframes} MODP(t)}{Nframes} \tag{4.10}$$

where: $Nframes$ is the total number of frames.

All the experiments were performed on a PC with an Intel (R) core (TM) i7-8565U CPU @ 1.8 GHz, 16 GB of RAM, using Matlab 2020a.

4.4.5 *Threshold definition*

Noise-induced entropy peaks need to be distinguished from actual objects-related entropy ones. A threshold-based decision is set in order to minimize the false positive rate of the proposed detector. Practically, we attempt to empirically identify an entropy threshold that allows the distinction of useful signal from noise.

To guarantee the most possible generalization, the OLIMP dataset has been randomly partitioned into two different subsets: reference and testing dataset. On that account, the threshold is determined related to the reference set, and the method is evaluated on the testing samples using the defined parameter. This evaluation methodology is coherent with state-of-the-art obstacle detection evaluation methodology[38]. To define the suitable value that ensures an accurate and reliable radar-based obstacle detection, the threshold has been explored based P, R and F1-score metrics. figure 4.7 shows the exploration results.

We selected the threshold that maximizes the F1-score. This choice is justified since F1-score is a measure that achieves a trade-off between P and R and yields to an accurate reliable system. The highest F1 occurs at the threshold value of **0.0153**, which is thereby selected as the empirical threshold.

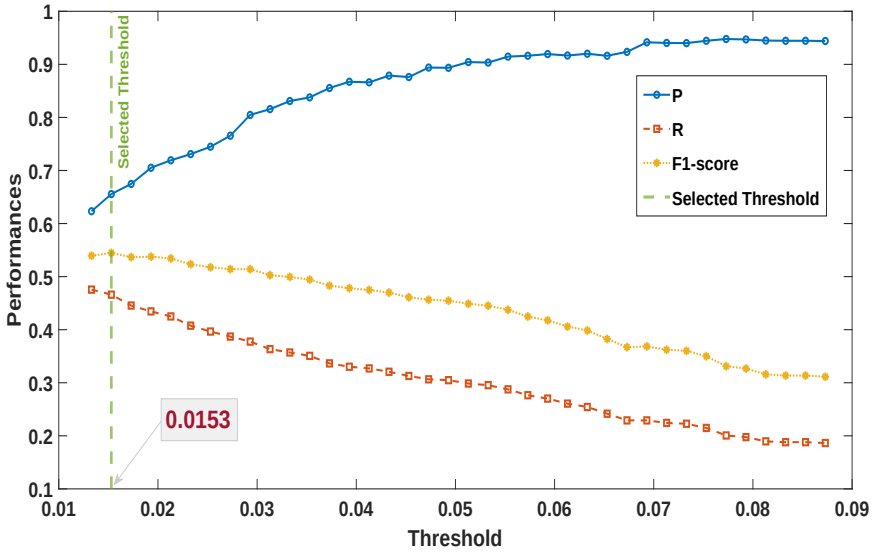


Figure 4.7: Threshold definition based on exhaustive space exploration. The threshold corresponds to the highest F1-score.

4.4.6 *Results*

To validate the efficiency of the proposed method, a comparative benchmarking is elaborated. The considered targets in this comparative study are: Pedestrian, Cyclist, Car and Tram. The obtained results are illustrated in figure 4.8 and figure 4.9. It can be seen from the figure that the proposed approach provides higher detection performances compared to HOS, CA-CFAR and [141] in terms of P, R and F1 score.

Since we are using real-world conditions benchmark, the CA-CFAR achieves the lowest performance. Its performance is degraded when challenged with the presence of multiple targets. HOS-based detector enhances the accuracy detection since among its characteristics it is able to suppress noise in the received signal. Nevertheless, HOS fails when a target is located further than another since its signature amplitude will be lower and the obstacle is consequently ignored.

To visualize this behavior, figure 4.10 gives an illustration of the aforementioned limitations for multiple targets detection. The figure shows that while HOS and CA-CFAR fail to accurately detect all the real targets presented in the UWB signal by generating either false positives or missdetections (false negatives), our method can determine the right targets' positions.

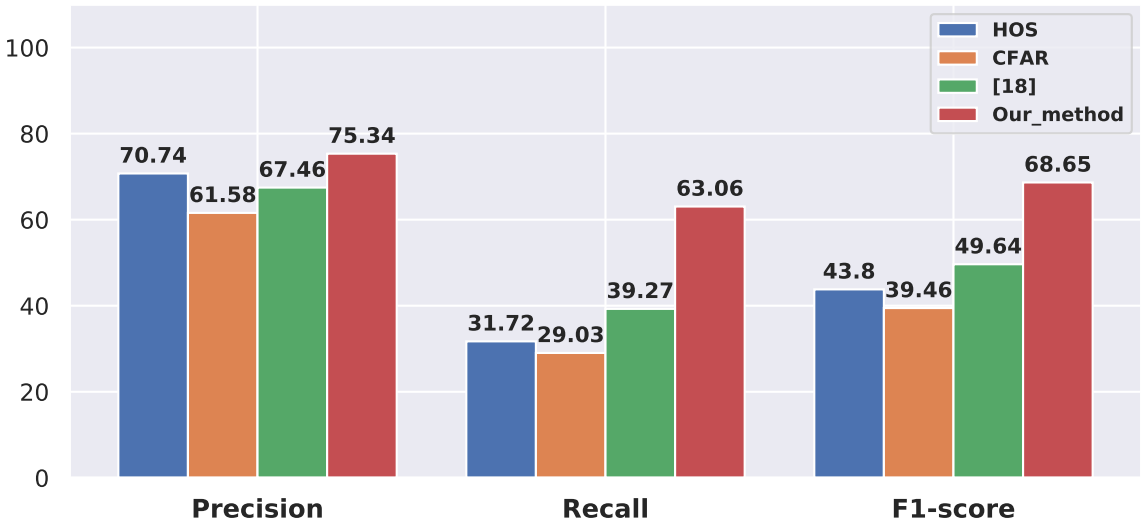


Figure 4.8: Experimental results: Precision, Recall and F1-score using HOS, CA-CFAR, [141] and our method.

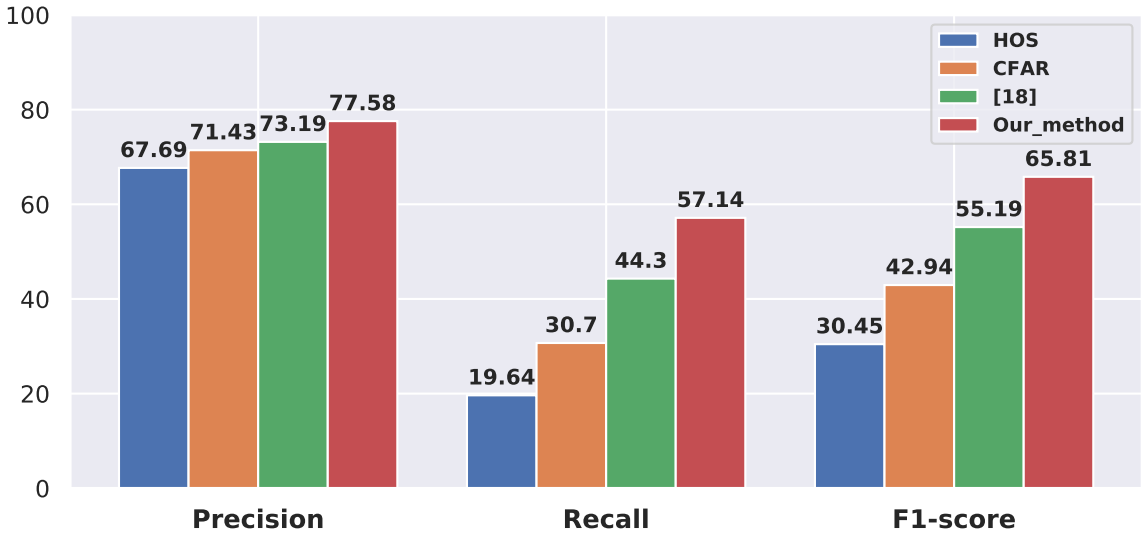


Figure 4.9: Experimental results for pedestrian detection using HOS, CA-CFAR, [141] and our method.

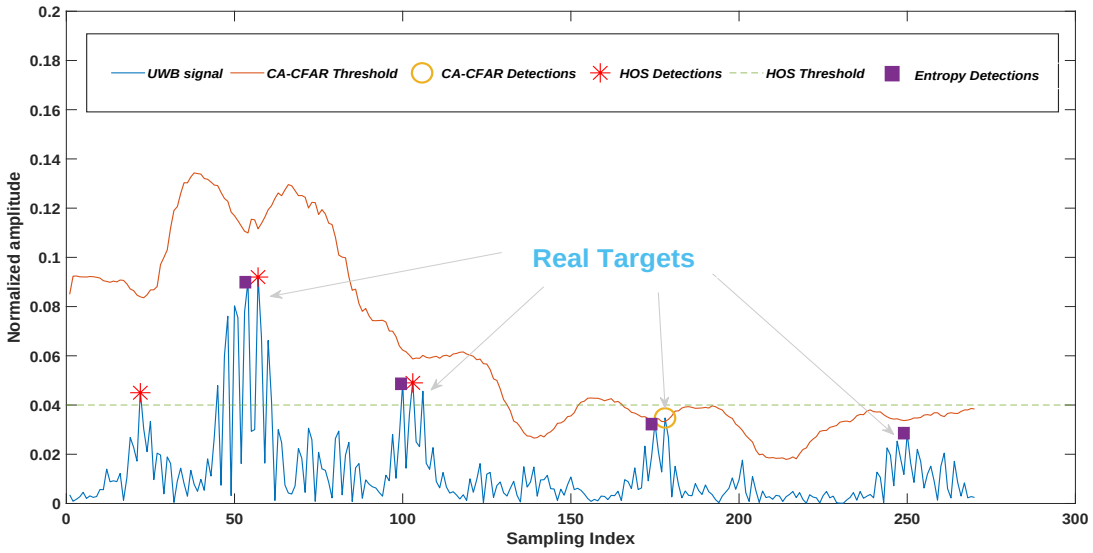


Figure 4.10: Illustration of multiple-target detection limitations using HOS, CA-CFAR and Entropy detectors

Due to physical characteristics, pedestrian's wave reflection amplitude is generally lower than other objects' reflections (especially metallic), and it gets further attenuated when moving away from the radar. This makes pedestrian detection challenging for HOS, CA-CFAR and [141]. Thus, we reported the performance related to pedestrian detection using the implemented detectors in figure 4.9. As depicted from results, even though the reflection intensity of a human body is low, the proposed method reports the best overall performance for pedestrian detection, and with even higher relative improvement. For example, while for all objects (figure 4.8), we achieve 6.11% higher precision than HOS, this improvement is up to 12.75% for pedestrian. In terms of recall, our approach provides an improvement of 49.7% for all objects and an increase of 65.63% for pedestrian detection compared with the same technique. These results show the robustness of our technique and can be explained by the fact that it is more concerned with entropy within the signal than the signal magnitude.

To further explore the effectiveness of entropy-based obstacle detection, we purpose to present results amplitude-wise. This experiment allows us to evaluate the robustness of the proposed technique comparatively both in terms of signal amplitude, but also in terms of the obstacle distance. In fact, the amplitude of a given reflected signal on an object decreases accordingly with the corresponding object distance from the radar.

Figures 4.11, 4.12 and 4.13 show respectively the precision, recall and F1-score of the different obstacle detection techniques as a function of the obstacle's amplitude (in the ground truth).

As reported in figure 4.11, HOS precision degrades proportionally with the target amplitude. Objects with low amplitude are not detected as this technique is sensible to high magnitude. CA-CFAR and [141] have practically balanced performance in terms of precision for all the amplitude ranges. Most of the state-of-the-art techniques show higher detection performance in high amplitude cases, i.e., close obstacles. Surprisingly, for our method, we noticed a counter-intuitive precision increase as the amplitude decreases. This observation will be explained later.

According to the presented results in figures 4.12 and 4.13, the performances of HOS, CA-CFAR and [141] degrade proportionally with decreasing the object amplitude, which is correlated with increasing the object distance. In fact, as the object is moving away from the radar field, its amplitude reflection attenuate and its signal amplitude becomes close to noise signals. Since these techniques depend particularly on the signal amplitude, several objects will not be detected, thereby resulting in false negatives, which explains the detection performance degradation (R). Nevertheless, the entropy-based method results increase remarkably as the object's amplitude attenuates. In low amplitude cases, our technique succeeds in limiting false negatives, thereby increasing recall. This observation is important since the proposed detector is robust even with challenging low amplitude targets. In the following we attempt to explain the observation of increasing detection performance with challenging signal situations.

We believe that this property is due to the increasing number of multipath components of the reflected signal, which increases the entropy within the corresponding signal. In fact, as illustrated in figure 4.14, the number of reflected signals goes higher with distant objects. While in the case of further objects the overall reflected signal amplitude is lower, magnitude-based techniques either loose in sensitivity or keep the same levels, our technique takes advantage from the differential entropy enhanced by the reflected multipath components. This observation is obviously limited by the radar range.

Furthermore, the MODP results related to the implemented detectors are reported in Table 4.3. The presented performances show that our detector can correctly detect targets position more than the other techniques with a MODP that reaches 0.42.

Finally, in terms of complexity, the execution time to detect obstacles in a received signal using the implemented detectors are reported in Table 4.3. From the results, we can observe that our proposed detector has the lowest execution time.

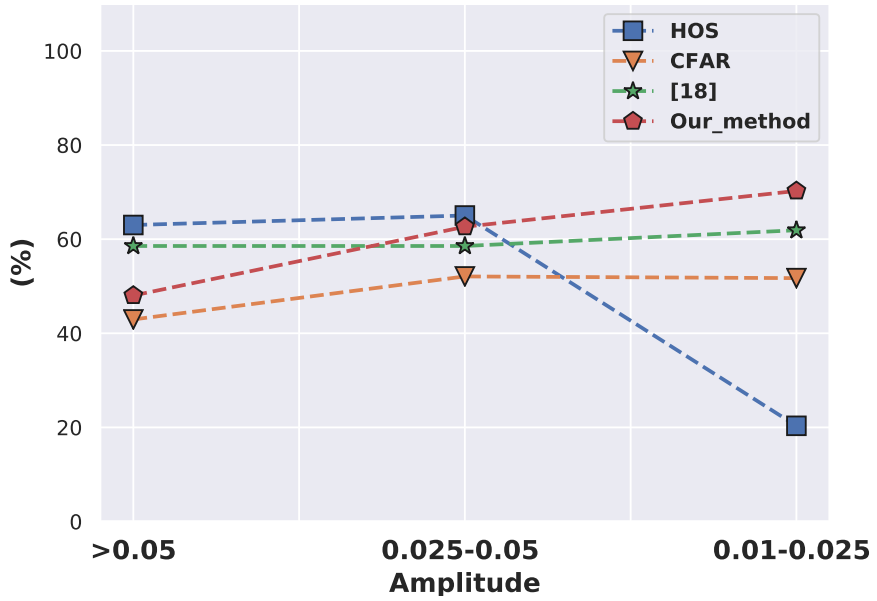


Figure 4.11: Precision results over varying target amplitude in terms of P, R and F1-score using HOS, CA-CFAR, [141] and our method.

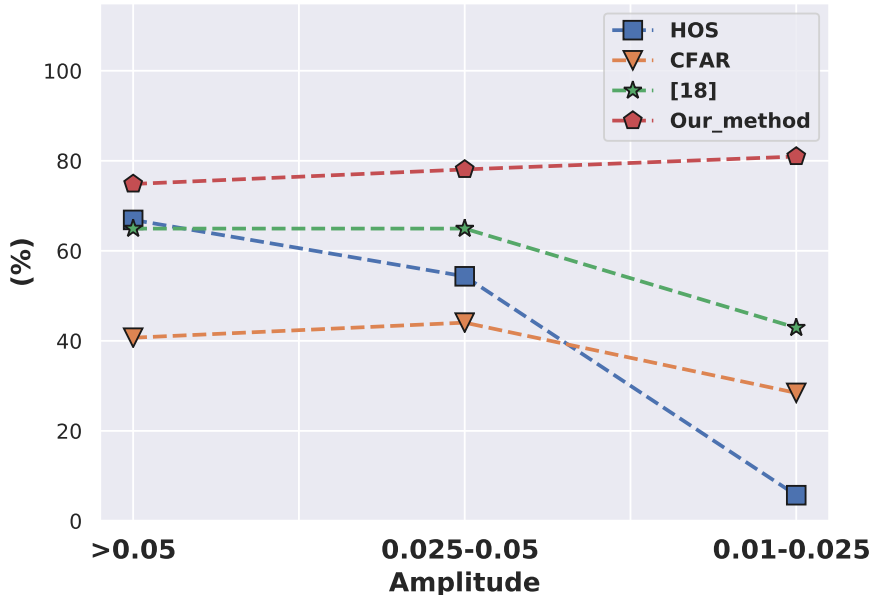


Figure 4.12: Recall results over varying target amplitude in terms of P, R and F1-score using HOS, CA-CFAR, [141] and our method.

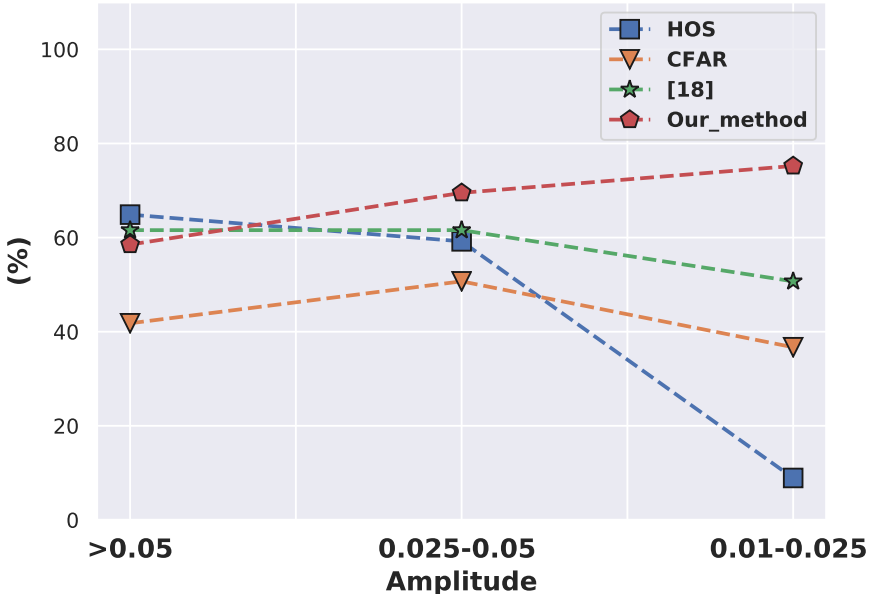


Figure 4.13: F1-score results over varying target amplitude in terms of P, R and F1-score using HOS, CA-CFAR, [141] and our method.

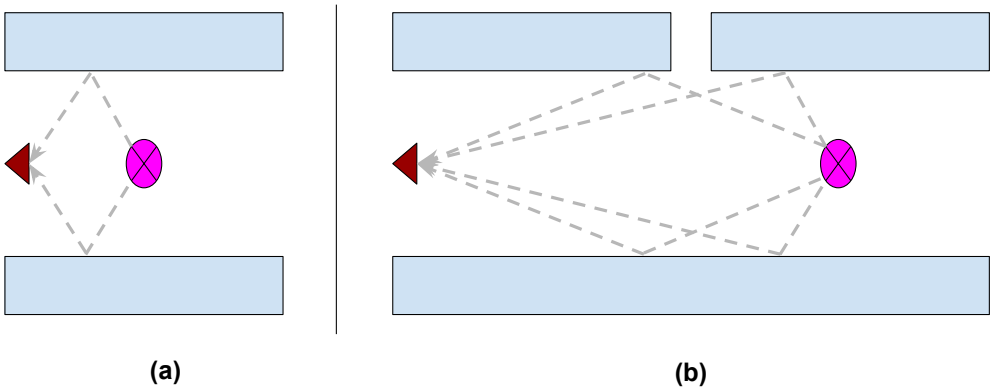


Figure 4.14: Illustration of multipath component vs distance: (a) close object with high amplitude, (b) object in further location with low amplitude.

Table 4.3: MODP and execution time results

	HOS	CFAR	[17]	Our Method
MODP	0.22	0.17	0.24	**0.42**
Execution time(s)	1.34	1.41	1.52	**1.26**

4.4.7 *Discussion*

An ITS is a safety-critical domain where achieving robust automatic environment perception is a challenging keystone. While most of the environment perception systems are based on camera and lidar sensors, UWB radars show interesting characteristics such as robustness to weather and luminosity challenges. This aspect could be complementary to other modalities and promising to enhance computer vision reliability.

In this work, an entropy-based ROI identification approach for UWB radar detector is proposed. The proposed technique is designed to identify relevant information in the received signal, hence, to differentiate a real object from noise based on their respective vehiculated entropy.

In fact, the entropy-based segmentation method for UWB radar signals technique exploits signal entropy instead of amplitude to localize useful parts of the signal, thereby detecting obstacles. A threshold based on the maximization of the F1-score is determined.

Our results show an overall improvement of detection performance using our technique compared with related work. In a detailed look at the results we made a surprising property of our technique that consists of an increase in detection robustness with lower amplitudes, and consequently further obstacles from the sensor. Although above the considered objects' distance, the method will be limited as Signal-to-noise ratio (SNR) will be very low, in our case we are interested in detecting near-range obstacles. For this reason, even with low amplitude, the entropy-based method can differentiate between noise and real target. This aspect is due to the wireless signal propagation multipath components on the signal entropy. On the other hand, there is room for improvement.

4.5 LSTM-BASED DETECTOR

4.5.1 *Background*

As we mentioned before, one of the fundamental challenges that could occur is distinguishing the real target's signature from noise within UWB signals. For this purpose, various studies have been conducted by proposing UWB-based detectors. Even though the interest of processing UWB data via deep learning techniques is growing, there is no work that treats deep with UWB signals for outdoor environment perception to the best of our knowledge.
In this section, we propose the first framework that exploits UWB signals with LSTM network in an outdoor environment involving complex urban driving situations for multi-target detection. As the UWB received signal is a time-series data, the exploitation of a RNNs is suitable to exhibit the temporal dependencies. Therefore, the main intention of the developed network is to discriminate the real target from noise within an UWB received signal. A comparison between our proposed approach and the state-of-the-art techniques based on expanded experiments using OLIMP is detailed in the following.

4.5.2 *Proposed LSTM-based detector*

In this section, we present firstly the background of the RNNs and LSTM networks, followed by a description of the proposed approach.

4.5.3 *RNNs for sequential data*

The independence among the data samples is one of the fundamental assumption for neural networks. Nevertheless, this assumption does not deal with sequential data as speech, video, time series, etc. The individual elements of this data exhibit dependency across time. Besides, the neural networks treat every data sample independently and thereby suffer the loss of benefiting from the exploitation of the sequential information. In addition, another disadvantage of the employment of neural networks is that they are not able to handle variable length sequences. For various domains like language translation or speech modeling, the sequences vary in length.

For these reasons, RNNs are introduced as a type of neural networks that are suitable for processing sequential data. In fact, this type of network processes the input sequence one element at a time and maintains a hidden state vector that serves as a memory for past information. RNNs learn to selectively conserve relevant information to capture dependencies within several time steps.

The architecture of RNN is shown in figure 4.15. The RNN has a feedback connection that connects the hidden neurons across time. At time t, it receives as input the current sequence's element x_t as input and the s_{t-1} the hidden state extracted from the previous time step. Afterwards, the hidden state is updated to s_t and h_t is calculated (the output of the network). U is the weight matrix that links the input and the hidden layers similar to a conventional neural network. The weight matrix related to the recurrent transition is presented by W, it links a hidden state to the next. V presents the weight matrix for hidden to the output transition.

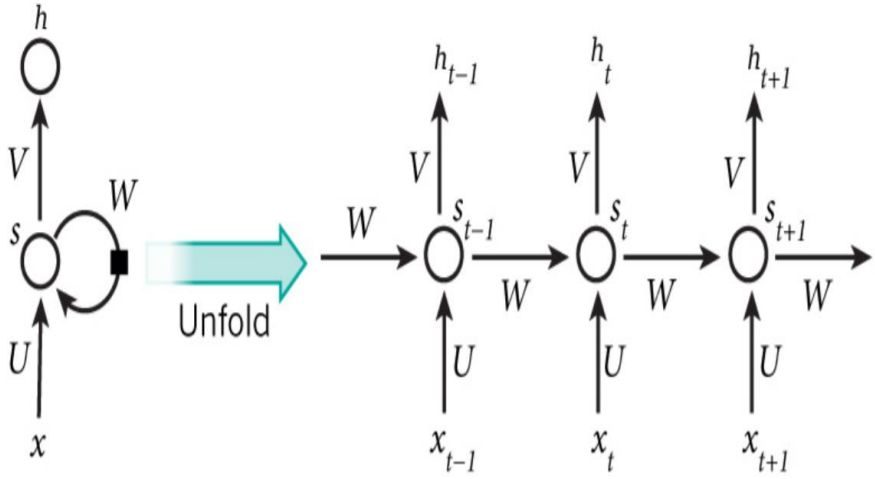


Figure 4.15: A standard RNN architecture. The left side of the figure represents a standard RNN [157].

The RNN is trained using the back-propagation through time (BPTT) [175] in order to learn long-range dependencies across long intervals. However, in practice training RNNs is a difficult process [10]. In fact, training the RNN with BPTT requires back-propagating the error gradients through several steps. According to the RNN shown in figure 4.15, the recurrent edge has the same weight in each time step. Hence, back-propagating the error implicates multiplying the error gradient together with the same value repeatedly. This causes that the gradient to either decay to zero or become too large with respect

to the layers' number [67]. These aforementioned problems are referred to vanishing gradients and exploding gradients respectively.

Consequently, numerous methods have been proposed to deal with the problems of learning long-term dependencies when training RNNs as the gradient clipping method that has proven to be effective to the exploding gradient problem [118]. Subsequently, the LSTM network has been introduced to overcome the vanishing problem due to the replacement of an ordinary neuron by a complex architecture entitled the LSTM unit [68]. The LSTM network has become the popular variant of the RNNs.

4.5.4 *LSTM background*

The **LSTM** network is a special architecture of the artificial RNNs developed in 1997 by Hochreiter and Schmidhuber [68]. It has been introduced to avoid the issues that occur when modeling long-term dependencies with RNN such as the vanishing or the exploding gradient problems. Therefore, the LSTM network is able to learn short-time as well as long-term dependencies. It is especially known by its effectiveness to treat time-series data [2]. In other words, the LSTM network is able to model the temporal changes in a series of data owing to its memory units and recurrent architecture. The LSTM units are connected sequentially. Each LSTM cell includes its own memory with three gates: the input, the output and the forget gates. These gates are responsible of protecting and controlling the flow of information through the cell. Otherwise, they decide which information has to be forgetting or reminded. These gates are detailed in the following:

- **The input gate**: It can allow the incoming signal to block the state of the memory cell or to change it.
- **The output gate**: It can authorize the state of the memory cell modify the other neurons or prevent it.
- **The forget gate**: It can let the cell to forget or remember its previous state, as required.

An LSTM cell can be expressed by Equation 4.11 and illustrated in figure 4.16.

$$\left[\begin{array}{l} I_t = \sigma(W_I[H_{t-1}, X_t] + b_I) \\ F_t = \sigma(W_F[H_{t-1}, X_t] + b_F) \\ O_t = \sigma(W_O[H_{t-1}, X_t] + b_O) \\ C_t = F_t * C_{t-1} + I_t * tanh(W_C[H_{t-1}, X_t] + b_C) \\ H_t = O_t * tanh(C_t) \\ Y_t = softmax(W_Y H_t + b_Y) \end{array}\right] \tag{4.11}$$

where:

- X=$\{$X(1),X(2),..,X(N)$\}$ is a an input sequence, where N is the length of the time series sequence.
- I_t, F_t, O_t and C_t are respectively the input, forget gate, the output gates and the memory cell state.
- H_t is the cell output and Y_t is the final output.
- $tanh$ and σ are respectively the hyperbolic tangent and the logistic sigmoid activation functions.
- W and b are respectively the input weights associating the LSTM cell to the inputs, and the bias vectors.

In fact, if a new input comes and the input gate I_t is activated, a new information will be added to the cell. Moreover, if the forget gate F_t was activated, the past cell status C_{t-1} could be forgotten. The output gate O_t controls either the last cell output C_t propagated into the final state H_t or not. The Nonlinear sigmoid $\sigma = (1 - e^{-1})^{-1}$ outputs values between zero and one, zero indicates that "let nothing through," while one means "let everything through!". Thus, the LSTM architecture utilizes the memory cells to use and store information, to identify the long-range temporal relations [111].

4.5.5 *Proposed UWB-based system for obstacle detection*

From the one side, despite the fact that UWB reflected signal incorporates rich information, the discrimination between the object's signature from noise is a fundamental

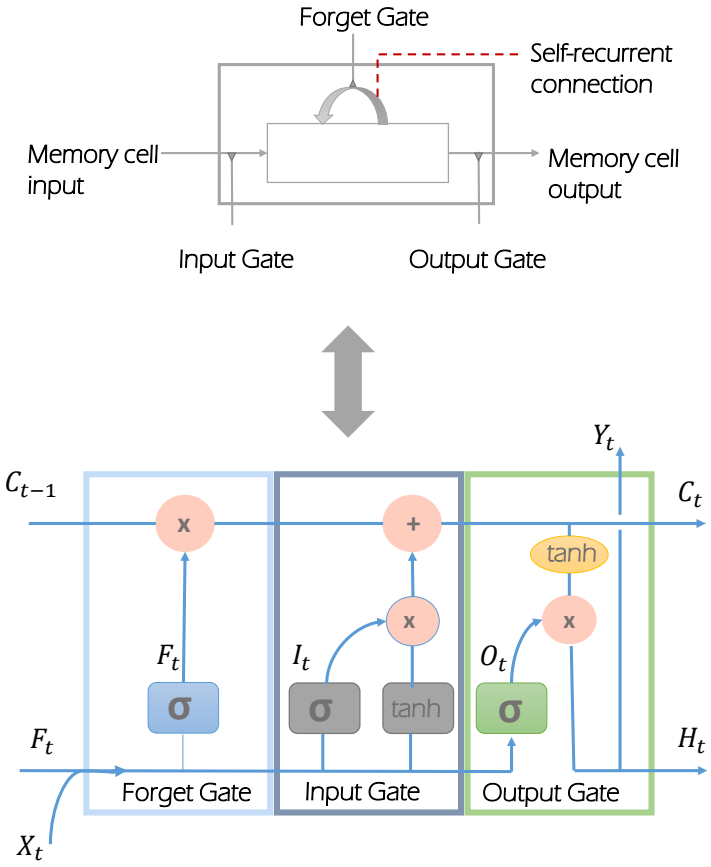


Figure 4.16: Architecture of LSTM unit

challenge. In fact, as we already mentioned that the UWB received signal $r(t)$ can be modulated following Equation 4.12.

$$r(t) = s(t) + n(t) \tag{4.12}$$

where $n(t)$ is the noise of the transmission channel of UWB radar and $s(t)$ is the received echo.

Therefore, based on the UWB property that indicates that each obstacle has its own signature, noise also should be different.

From the other side, the target's wave reflection amplitude is generally changing over time due to physical characteristics (material, shape, size, etc.). Otherwise, the received echo is a function of time, and also a function of the obstacle distance from the sensor. This time-distance relation is expressed by Equation 4.1.

Hence, on account of the temporal changes presented in the UWB signal we adopt the idea of employing LSTM network. This choice is made on account of the fact that this type of RNN is able to recognize and synthesize the dynamics variations within the UWB received echo. Thereby, in this work we propose the first framework that exploits LSTM network with UWB signals for distinguishing obstacles from noise in a vehicle environment perception context.

The proposed framework is explained in Algorithm 2 and illustrated in figure 4.17. Primarily, the received echo is split into time series sequences $[x_t]$. These sequences present the useful regions which contain the real targets (illustrated by the orange windows in figure 4.17) and noise partitions (depicted by green rectangles in figure 4.17). For data variety, the noise parts are randomly selected from the UWB signal. In fact, the window's size ($Wsig$) is empirically set according to the objects' signature length. Afterwards, features from the time-frequency domain are extracted from the defined regions. Thus, the discrete wavelet transform (DWT) is utilized. We extract four features from the approximation coefficients (Ca) and the detail ones (Cd) for each sequence. Subsequently, the fattened 1D descriptor vector feeds the LSTM network. Finally, the output of the LSTM is linked to the fully connected layer of size two followed by a Softmax layer, and a classification layer.

To conclude, the proposed detector analyzes the temporal changes within the UWB signal via learning the extracted time-frequency features that highly present the signal characteristics.

Algorithmus 2 : LSTM-based algorithm for obstacle detection using UWB signals

Data : Received signal $r = [r_t]$, Signature window length: $Wsig$; $Lstm_{options}$; Position: pos

Output : Class

// Split time series sequence
for $t \leftarrow pos$ **to** $length(Wsig)$ **do**
    $x_t = r(pos, t + pos)$;
end for
// Extract features using Discrete wavelet transform
$Fvector = []$
$[Ca, Cd] = \texttt{DWT}(x_t)$
$F1 = mean(Ca); F2 = std(Ca); F3 = min(Cd); F4 = rms(Cd)$
$Fvector = append(Fvector, F1, F2, F3, F4)$
// Feed the LSTM-based network
$Output_{lstm} = Lstm(Fvector, Lstm_{options})$
$OutputFCL = Fullyconnectedlayer(Output_{lstm})$
$Class = Softmax(OutputFCL)$
return $(Class)$

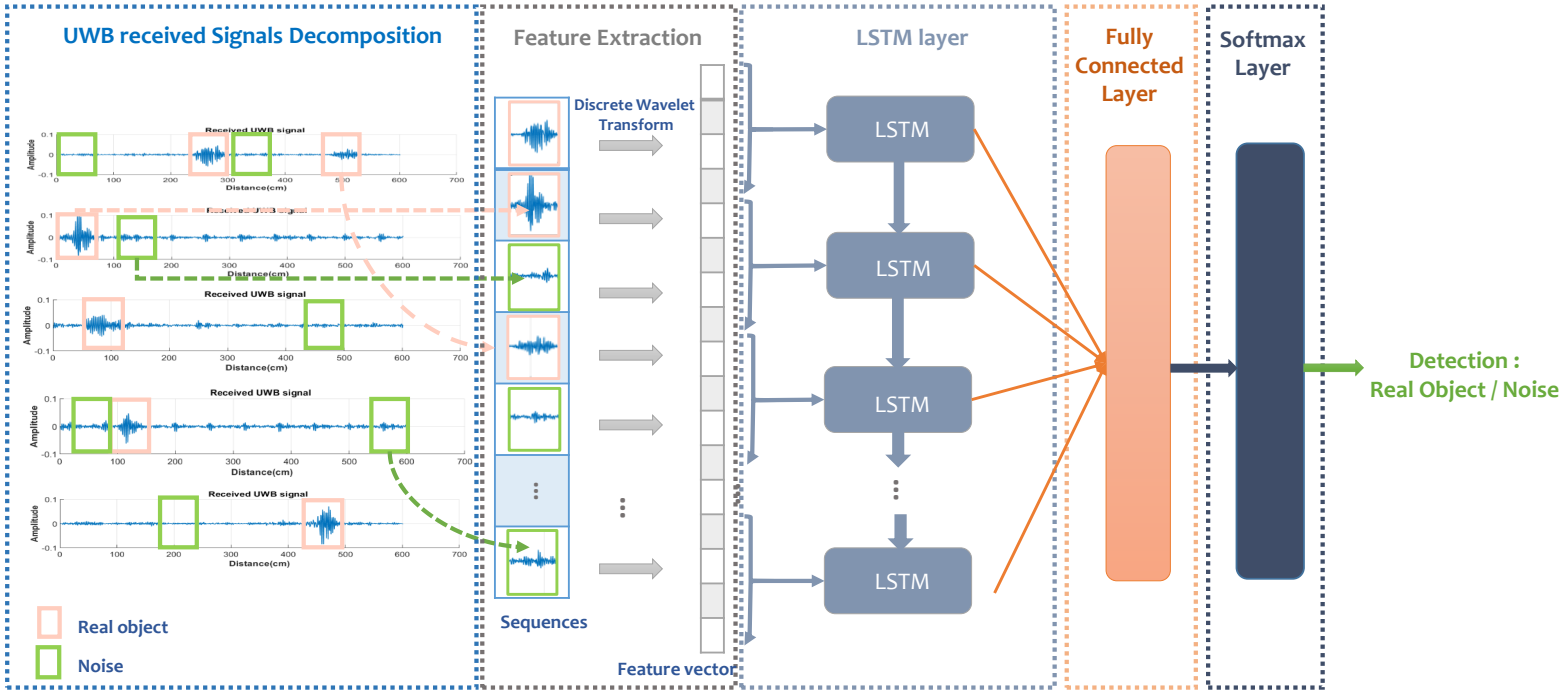


Figure 4.17: Proposed LSTM-based framework using UWB signals

4.5.6 *Experimental setup*

To highlight the efficiency of the proposed architecture, we compare it to the implemented state-of-the art techniques: CFAR, HOS and the work in [141]. The experimental results are evaluated using the P, R and the F1-score metrics.

The details of the implementation are illustrated below:

- **Network architecture :** We exploit the unidirectional network with 100 LSTM hidden units. The number of epochs and mini batch size during experiments are set to 100 and 64 instances respectively
- **Feature extraction :** To extract features using the discrete wavelet transform, the Daubechies wavelet is employed to extract the features from the time series sequences.
- **Dataset :** We conducted our experiments on a variety of the OLIMP urban driving scenarios.
- **Training :** For our experiments, we use $2/3$ of the data for training step and $1/3$ for the test process following the OLIMP protocol.
  Moreover, we utilized the Adam Optimizer [14]. The initial learning rate is set to 0.001.
- **Comparative techniques :** For HOS we take advantage of the 4^{th} order cumulant that relies on Tuganit4 algorithm. Concerning CFAR, CA-CFAR detector is considered with automatic threshold. These techniques are already explained in Section 4.4.4.

It shall be mentioned that for training process, The optimal set of parameters has been selected based on preliminary tests.

4.5.7 *Results*

The experimental results concerning the comparative study are summarized in figure 4.18. In fact, it can be seen from the figure that the obtained results show that our deep learning-based method achieves the highest performance. Our proposed method

outperforms significantly the considered traditional detectors particularly in terms of recall and precision.

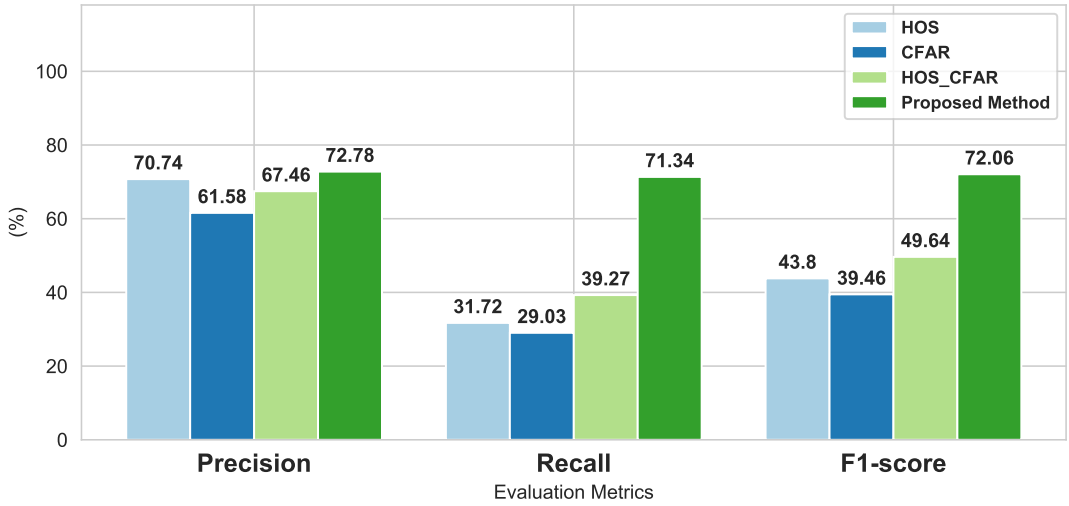


Figure 4.18: Experimental results using HOS, CFAR, the work in [141] and our method

The performance of CA-CFAR, HOS and their combination in the work of [141] depend essentially on the definition of the threshold parameter. A higher threshold generates more false negatives, however, the opposite case produces additional false positives. Furthermore, the object detection rate using the aforementioned techniques rely on the object's amplitude. In low magnitude cases, which means that the object is moving away from the radar, the target can't be detected and it is considered as noise.

The LSTM-based method can distinguish noise from real targets thanks to the relevant features that are extracted from the time-frequency domain, and by learning the temporal relationship between the data sequences. In fact, the time-frequency features lead to a high-performing as they can represent well the signal's characteristics. It shall be mentioned that, some missclassifications are still present due the challenges related to the interclass similarity of the obstacle's signature and the noise signal in cases where the object's signature has a low amplitude.

In terms of complexity, table 4.4 depicts a comparison of the execution time of our method with the state-of-the-art techniques. As it can be seen from the results, our architecture has the highest execution time, though, advanced hardware resources are to be deployed to acquire faster detection time.

Table 4.4: Execution time results

	HOS	CA-CFAR	[141]	Proposed method
Execution time (s)	1.34	1.41	1.52	2.01

4.6 DISCUSSION

In this chapter, we tackled the problem of differentiating real objects from noise for the environment perception purpose using UWB radar. For this purpose, we proposed to explore an entropy-based method and a deep-based framework. We compared the proposed approaches to the well-known techniques HOS, CA-CFAR and [141] and experiments were conducted on our previously developed dataset OLIMP.

Even though the main aim is to discriminate the useful portions within the signal from noise and to detect the obstacle, the two proposed detectors treat the dilemma differently. On the one hand, the entropy-based method exploits the signal's complexity instead of the amplitude which the state-of-the art techniques rely on it. This leads us to prove that our entropy-detector remains robust even with lower amplitude cases. As a future work, it should be pointed out that an adaptive thresholding could be considered in our case.

On the other hand, the LSTM-based detector aims to localise the real obstacles by learning and synthesizing the temporal variations within the UWB received data sequences. This detector achieves higher results in terms of recall, but not in terms of precision compared to our proposed entropy-based detector. In terms of complexity, the execution time of the entropy-based method is faster to detect the obstacles compared with the LSTM-based as it is a time consuming (deep-based method) and all the implemented methods.

While the LSTM-based network has proven to be particularly powerful to solve noise and real target distinguishing problems, the entropy characterises the UWB signal, therefore, a hybrid approach that combines these methods could be interesting to compensate their limitations and obtain a better detector.

4.7 CONCLUSION

In this chapter, we focused on the segmentation of 1D UWB signals. Thus, we reviewed the existing methods to detect obstacle using UWB radar. Afterwards, we proposed an

entropy-based detector based on a theoretical study related to the Shannon entropy. A comparative study has been conducted to evaluate our entropy method by employing HOS, CA-CFAR and [141] techniques. In addition, as a deep-based method, we proposed the first framework that exploits LSTM network to distinguish real target from noise using UWB radar.

According to the obtained results that surpass the state-of-the-art techniques, the proposed systems constitute an important step towards distinguishing real obstacles from noise. As future work, we believe that an adaptive hybrid segmentation technique could be interesting and may achieve high detection performance.

5

CONCLUSIONS AND PERSPECTIVES

Contents

5.0.1 *Conclusions and contributions*

Environment perception is one of the key challenges in automated driving applications especially in the quest for higher degrees of automation. These systems represent the the forthcoming of transportation. In fact, early ADASs have been developed and improved over many generations thanks to the better performance of the sensors and their enhancement of processing algorithms. In the recent years, sensor data fusion becomes a key aspect in developing such systems to describe the complete vehicle environment optimally and efficiently. Therefore, since the level of automation increases and ADASs become complicated, reliable environment perception is required. Moreover, developing accurate and reliable systems will always be addressed particularly for next-generation automated systems as they guarantee safety.

To address this challenge, in this thesis, we have focused on environment perception. Particularly, we have tackled the problem of detecting multiple objects using data fusion or a unique sensor. This thesis integrates four principal contributions.

Firstly, we have reviewed the environment perception state-of-the-art for intelligent and autonomous vehicles, specifically the object detection task which includes the relevant sensors, the data fusion methods and the challenges. Afterwards, by reviewing the existing public multi-modal environment perception databases, we have introduced OLIMP dataset. It is the first synchronized dataset that includes these four modalities: images, UWB radar signals, narrow band data streams and acoustic data.

We have also presented a new fusion framework that combines data acquired from different sensors used in our dataset to achieve better performances for the obstacle detection task. Various levels of fusion are exploited and promising results are found. These results undoubtedly allow us to conclude that multimodality is indeed to guarantee an accurate environment perception.

Afterwards, we have tackled the problem of multi-obstacle detection in short-range settings using the UWB radar as it provides rich information about the vehicle's surroundings. Thus, two detectors are proposed.

The first detector is an entropy-based segmentation approach. It exploits the signal entropy instead of the amplitude to localize the useful UWB parts. Based on exhaustive experiments and a comparative study with the state-of-the-art techniques our method remains robust even with challenging low amplitude signals. Furthermore, our detector's

performance is improved when the obstacle is far from the sensor. This observation overcomes the fundamental challenge in related techniques.

The second detector is a deep learning-based framework. It is the first framework that exploits LSTM with UWB signals for multi obstacle detection in an outdoor complex environment. By learning the temporal dependencies within UWB series of data, the LSTM-based detector outperforms the conventional techniques.

Finally, based on the aforementioned contributions made in this thesis dealing with the problem of object detection, we conclude that for sensor data fusion architectures there is not a dominant level of fusion. In fact, it depends particularly on the targeted ADAS applications, the employed sensors and the environment. In addition, in urban environments, the environment perception task is more challenging as they are very dense and it is difficult to separate objects near each other.

5.0.2 *Perspectives*

Despite these advances made through our contributions, the complete environment perception remains an open topic of research and several improvements can be envisaged. principal limitations to be considered and the perspectives are presented as follows.

- **Sensor fusion**

  The proposed fusion framework have shown that data fusion at different levels offer a higher performance than employing a unique sensor. The proposed fusion framework is limited because of its sequential aspect. We believe that this could be improved using advanced parallel fusion systems. This will be investigated in future work.

  Despite the fact that the suggested fusion detection method deals with the limitations of each modality, only the detection of false positives is considered. Incorporating the cases of false negatives in the fusion framework can be explored as future work to obtain a more reliable object detection system.

  Furthermore, the development of a new deep learning-based architecture that fuses narrow-band data streams and images is our current field of research.

- **Entropy-based detector**

The entropy-based segmentation method for UWB radar signals exploits the signal entropy. It relies on the detection of a threshold that allows the distinction of useful signals from noise. It should be pointed out that adaptive thresholding could be considered in our case.

Even with a low amplitude, the entropy-based method can differentiate between noise and real targets. However, we believe that an adaptive hybrid segmentation technique could be interesting and may achieve a high detection performance.

- **LSTM-based detector**

  We have shown that analyzing the temporal data changes among UWB signals for the object detection process provide higher performances by exploiting RNNs. Promising results are found using the LSTM-based detector which outperforms the performances of the related techniques. It shall be mentioned that some miss-classifications are still present. Accordingly, further research can be expanded to extract deep features to tackle this dilemma. For this reason, the employment of CNNs to extract features from the UWB signals can be investigated.

  From another perspective, an hybrid approach that combines the entropy-based detector and the LSTM one could be interesting to compensate limitations and obtain a better detector.

Part I

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Shahzad Ahmed and Sung Ho Cho. "Hand Gesture Recognition Using an IR-UWB Radar with an Inception Module-Based Classifier." In: *Sensors* 20.2 (2020), p. 564.

[2] Safa Ameur, Anouar Ben Khalifa, and Med Salim Bouhlel. "A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with Leap Motion." In: *Entertainment Computing* 35 (2020), p. 100373.

[3] Moeness G Amin and Baris Erol. "Understanding deep neural networks performance for radar-based human motion recognition." In: *2018 IEEE Radar Conference (RadarConf18)*. IEEE. 2018, pp. 1461–1465.

[4] Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. "A survey on 3d object detection methods for autonomous driving applications." In: *IEEE Transactions on Intelligent Transportation Systems* 20.10 (2019), pp. 3782–3795.

[5] *Axis*. https://www.axis.com/products/axis-m1113. 2018.

[6] Mark Bagnoli and Ted Bergstrom. "Log-Concave Probability and Its Applications." In: *Economic Theory* 26.2 (2005), pp. 445–469. ISSN: 09382259, 14320479.

[7] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset." In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 6433–6438.

[8] Behnam Behroozpour, Phillip AM Sandborn, Ming C Wu, and Bernhard E Boser. "Lidar system architectures and circuits." In: *IEEE Communications Magazine* 55.10 (2017), pp. 135–142.

[9] Jorge Beltrán, Carlos Guindel, Francisco Miguel Moreno, Daniel Cruzado, Fernando Garcia, and Arturo De La Escalera. "Birdnet: a 3d object detection framework from lidar information." In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2018, pp. 3517–3523.

[10] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. “Learning long-term dependencies with gradient descent is difficult.” In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166.

[11] Mario Bijelic, Fahim Mannan, Tobias Gruber, Werner Ritter, Klaus Dietmayer, and Felix Heide. “Seeing through fog without seeing fog: Deep sensor fusion in the absence of labeled training data.” In: *arXiv preprint arXiv:1902.08913* (2019).

[12] Stephen Blake. “OS-CFAR theory for multiple targets and nonuniform clutter.” In: *IEEE transactions on aerospace and electronic systems* 24.6 (1988), pp. 785–790.

[13] S. Bobkov and M. Madiman. “The Entropy Per Coordinate of a Random Vector is Highly Constrained Under Convexity Conditions.” In: *IEEE Transactions on Information Theory* 57.8 (2011), pp. 4940–4954.

[14] Sebastian Bock, Josef Goppold, and Martin Weiß. “An improvement of the convergence proof of the ADAM-Optimizer.” In: *arXiv preprint arXiv:1804.10587* (2018).

[15] Marta Borowska. “Entropy-based algorithms in the analysis of biomedical signals.” In: *Studies in Logic, Grammar and Rhetoric* 43.1 (2015), pp. 21–32.

[16] Marta Borowska. “Entropy-based algorithms in the analysis of biomedical signals.” In: *Studies in Logic, Grammar and Rhetoric* 43.1 (2015), pp. 21–32.

[17] Mokhtar Bouain, Denis Berdjag, Nizar Fakhfakh, and Rabie Ben Atitallah. “Multi-sensor fusion for obstacle detection and recognition: A belief-based approach.” In: *2018 21st International Conference on Information Fusion (FUSION)*. IEEE. 2018, pp. 1217–1224.

[18] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. “Segmentation and recognition using structure from motion point clouds.” In: *European conference on computer vision*. Springer. 2008, pp. 44–57.

[19] Travis D Bufler and Ram M Narayanan. “Radar classification of indoor targets using support vector machines.” In: *IET Radar, Sonar & Navigation* 10.8 (2016), pp. 1468–1476.

[20] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. “nuscenes: A multimodal dataset for autonomous driving.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 11621–11631.

[21] Long Cai, Xiaochuan Ma, Shefeng Yan, Chengpeng Hao, and Rongbo Wang. "Some analysis of fuzzy CAGO/SO CFAR detector in non-Gaussian background." In: *2010 2nd International Workshop on Intelligent Systems and Applications*. IEEE. 2010, pp. 1–4.

[22] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. "A unified multi-scale deep convolutional neural network for fast object detection." In: *European conference on computer vision*. Springer. 2016, pp. 354–370.

[23] Simon Chadwick, Will Maddetn, and Paul Newman. "Distant vehicle detection using radar and vision." In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 8311–8317.

[24] Sixian Chen, Chongyi Fan, Xiaotao Huang, and Chun Cao. "Low PRF Low Frequency Radar Sensor for Fall Detection by Using Deep Learning." In: *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*. IEEE. 2019, pp. 400–404.

[25] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. "Monocular 3d object detection for autonomous driving." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2147–2156.

[26] Yiping Chen, Jingkang Wang, Jonathan Li, Cewu Lu, Zhipeng Luo, Han Xue, and Cheng Wang. "Lidar-video driving dataset: Learning driving policies effectively." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5870–5878.

[27] Chia-Chin Chong, Fujio Watanabe, and Hiroshi Inamura. "Potential of UWB technology for the next generation wireless communications." In: *2006 IEEE Ninth International Symposium on Spread Spectrum Techniques and Applications*. IEEE. 2006, pp. 422–429.

[28] E Conte, M Longo, and M Lops. "Performance analysis of CA-CFAR in the presence of compound Gaussian clutter." In: *Electronics Letters* 24.13 (1988), pp. 782–783.

[29] *Continental*. https://www.conti-engineering.com/en-US/Industrial-Sensors/Sensors/. 2018.

[30] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. "The cityscapes dataset for semantic urban scene understanding." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3213–3223.

[31] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. "The cityscapes dataset for semantic urban scene understanding." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3213–3223.

[32] *Crash Google*. https://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash/. 2021.

[33] *Crash Hyundai*. https://jalopnik.com/this-is-how-bad-self-driving-cars-suck-in-the-rain-1666268433. 2021.

[34] *Crash Tesla*. https://money.cnn.com/2016/07/07/technology/tesla-liability-risk/index.html. 2021.

[35] *Crashes*. 2021.

[36] G Richard Curry. *Radar essentials: a concise handbook for radar design and performance*. The Institution of Engineering and Technology, 2012.

[37] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. "R-fcn: Object detection via region-based fully convolutional networks." In: *Advances in neural information processing systems* 29 (2016), pp. 379–387.

[38] Abdelkader Dairi, Fouzi Harrou, Mohamed Senouci, and Ying Sun. "Unsupervised obstacle detection in driving environments using deep-learning-based stereovision." In: *Robotics and Autonomous Systems* 100 (2018), pp. 287–301.

[39] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection." In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 886–893.

[40] Mark E Davis. "Ultra Wide Band Surveillance Radar." In: *2019 IEEE Radar Conference (RadarConf)*. IEEE. 2019, pp. 1–175.

[41] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. "Fast feature pyramids for object detection." In: *IEEE transactions on pattern analysis and machine intelligence* 36.8 (2014), pp. 1532–1545.

[42] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. "Integral channel features." In: (2009).

[43] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. "CARLA: An open urban driving simulator." In: *arXiv preprint arXiv:1711.03938* (2017).

[44] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks." In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 1355–1361.

[45] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "Density-based clustering algorithms for discovering clusters." In: *Kdd-96 Proceedings* 2 (1996), pp. 226–231.

[46] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. "Object detection with discriminatively trained part-based models." In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2009), pp. 1627–1645.

[47] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer. "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges." In: *IEEE Transactions on Intelligent Transportation Systems* (2020), pp. 1–20.

[48] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges." In: *IEEE Transactions on Intelligent Transportation Systems* (2020).

[49] Merzak Ferroukhi, Abdeldjalil Ouahabi, Mokhtar Attari, Yassine Habchi, and Abdelmalik Taleb-Ahmed. "Medical video coding based on 2nd-generation wavelets: Performance evaluation." In: *Electronics* 8.1 (2019), p. 88.

[50] Michael S Foster, John R Schott, and David W Messinger. "Spin-image target detection algorithm applied to low density 3D point clouds." In: *Journal of Applied Remote Sensing* 2.1 (2008), p. 023539.

[51] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. "Virtual worlds as proxy for multi-object tracking analysis." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4340–4349.

[52] Vladimir G Galushko. "Analysis of the CA CFAR algorithm as applied to detection of stationary Gaussian signals against a normal noise background." In: *2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW)*. IEEE. 2016, pp. 1–3.

[53] Jonah Gamba. "Radar Target Detection." In: *Radar Signal Processing for Autonomous Driving*. Springer, 2020, pp. 53–63.

[54] Yuan Gao, Xiao Ai, Yifei Wang, J Rarity, and Naim Dahnoun. "UV-disparity based obstacle detection with 3D camera and steerable filter." In: *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2011, pp. 957–962.

[55] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 3354–3361.

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

[57] 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*. 2014, pp. 580–587.

[58] Alejandro González, David Vázquez, Antonio M López, and Jaume Amores. "On-board object detection: Multicue, multimodal, and multiview random forest of local experts." In: *IEEE transactions on cybernetics* 47.11 (2016), pp. 3980–3990.

[59] Ian Gresham, Alan Jenkins, Robert Egri, Channabasappa Eswarappa, Noyan Kinayman, Nitin Jain, Richard Anderson, Frank Kolak, Ratana Wohlert, Shawn P Bawell, et al. "Ultra-wideband radar sensors for short-range vehicular applications." In: *IEEE transactions on microwave theory and techniques* 52.9 (2004), pp. 2105–2122.

[60] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection." In: *Information Fusion* 50 (2019), pp. 148–157.

[61] Junyao Guo, Unmesh Kurup, and Mohak Shah. "Is It Safe to Drive? An Overview of Factors, Metrics, and Datasets for Driveability Assessment in Autonomous Driving." In: *IEEE Transactions on Intelligent Transportation Systems* (2019).

[62] Joko Hariyono, Van-Dung Hoang, and Kang-Hyun Jo. "Moving object localization using optical flow for pedestrian detection from a moving vehicle." In: *The Scientific World Journal* 2014 (2014).

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Spatial pyramid pooling in deep convolutional networks for visual recognition." In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015), pp. 1904–1916.

[64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[65] Steffen Heuel and Hermann Rohling. "Pedestrian recognition based on 24 GHz radar sensors." In: *11-th International Radar Symposium*. IEEE. 2010, pp. 1–6.

[66] Steffen Heuel and Hermann Rohling. "Two-stage pedestrian classification in automotive radar systems." In: *2011 12th International Radar Symposium (IRS)*. IEEE. 2011, pp. 477–484.

[67] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. 2001.

[68] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[69] Thorsten Hoeser and Claudia Kuenzer. "Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends." In: *Remote Sensing* 12.10 (2020), p. 1667.

[70] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. "Speed/accuracy trade-offs for modern convolutional object detectors." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7310–7311.

[71] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. "The apolloscape open dataset for autonomous driving and its application." In: *IEEE transactions on pattern analysis and machine intelligence* 42.10 (2019), pp. 2702–2719.

[72] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. "Multispectral pedestrian detection: Benchmark dataset and baseline." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1037–1045.

[73] Paul Jaccard. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura." In: *Bull Soc Vaudoise Sci Nat* 37 (1901), pp. 547–579.

[74] I. Jegham, A. Ben Khalifa, I. Alouani, and M. A. Mahjoub. "Safe Driving : Driver Action Recognition using SURF Keypoints." In: *2018 30th International Conference on Microelectronics (ICM)*. 2018, pp. 60–63.

[75] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. "MDAD: A Multimodal and Multiview in-Vehicle Driver Action Dataset." In: *Computer Analysis of Images and Patterns*. Ed. by Mario Vento and Gennaro Percannella. Cham: Springer International Publishing, 2019, pp. 518–529. ISBN: 978-3-030-29888-3. DOI: 10.1007/978-3-030-29888-3_42.

[76] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. "Vision-based human action recognition: An overview and real world challenges." In: *Forensic Science International: Digital Investigation* 32 (2020), p. 200901.

[77] Changhui Jiang, Jichun Shen, Shuai Chen, Yuwei Chen, Di Liu, and Yuming Bo. "UWB NLOS/LOS Classification Using Deep Learning Method." In: *IEEE Communications Letters* 24.10 (2020), pp. 2226–2230.

[78] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. "A survey of deep learning-based object detection." In: *IEEE Access* 7 (2019), pp. 128837–128868.

[79] Yue Kang, Hang Yin, and Christian Berger. "Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments." In: *IEEE Transactions on Intelligent Vehicles* 4.2 (2019), pp. 171–185.

[80] Bushra Khalid, Asad Mansoor Khan, Muhammad Usman Akram, and Sherin Batool. "Person detection by fusion of visible and thermal images using convolutional neural network." In: *2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE)*. IEEE. 2019, pp. 143–148.

[81] Anouar Ben Khalifa, Ihsen Alouani, Mohamed Ali Mahjoub, and Najoua Essoukri Ben Amara. "Pedestrian detection using a moving camera: A novel framework for foreground detection." In: *Cognitive Systems Research* 60 (2020), pp. 77–96.

[82] Dae-Hyun Kim. "Lane Detection Method with Impulse Radio Ultra-Wideband Radar and Metal Lane Reflectors." In: *Sensors* 20.1 (2020), p. 324.

[83] Sangtae Kim, Seunghwan Lee, Seungho Doo, and Byonghyo Shim. "Moving target classification in automotive radar systems using convolutional recurrent neural networks." In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pp. 1482–1486.

[84] Woosuk Kim, Hyunwoong Cho, Jongseok Kim, Byungkwan Kim, and Seongwook Lee. "YOLO-Based Simultaneous Target Detection and Classification in Automotive FMCW Radar Systems." In: *Sensors* 20.10 (2020), p. 2897.

[85] Seok-Kap Ko and Byung-Tak Lee. "Object classification of UWB responses using S T-CNN." In: *2016 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE. 2016, pp. 794–796.

[86] Sampo Kuutti, Saber Fallah, Konstantinos Katsaros, Mehrdad Dianati, Francis Mccullough, and Alexandros Mouzakitis. "A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications." In: *IEEE Internet of Things Journal* 5.2 (2018), pp. 829–846.

[87] Jean Lahoud and Bernard Ghanem. "2d-driven 3d object detection in rgb-d images." In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4622–4630.

[88] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. "Pointpillars: Fast encoders for object detection from point clouds." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12697–12705.

[89] *Level of automation*. https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic. 2020.

[90] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. "Illumination-aware faster R-CNN for robust multispectral pedestrian detection." In: *Pattern Recognition* 85 (2019), pp. 161–171.

[91] Q. Liang, B. Zhang, and X. Wu. "UWB radar for target detection: DCT versus matched filter approaches." In: *2012 IEEE Globecom Workshops*. 2012, pp. 1435–1439.

[92] Sohee Lim, Jaehoon Jung, Seong-Cheol Kim, and Seongwook Lee. "Deep Neural Network-Based In-Vehicle People Localization Using Ultra-Wideband Radar." In: *IEEE Access* (2020).

[93] 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*. Springer. 2016, pp. 21–37.

[94] Y Liu, Z Wang, Z Zhang, and H Xu. "Reversing obstacle detection based on binocular vision image." In: *J. Chongqing Jiaotong Univ. Natural Sci.* 37.3 (2018), pp. 92–98.

[95] Jakob Lombacher, Markus Hahn, Jürgen Dickmann, and Christian Wöhler. "Potential of radar for static object classification using deep learning methods." In: *2016 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*. IEEE. 2016, pp. 1–4.

[96] Jakob Lombacher, Markus Hahn, Jürgen Dickmann, and Christian Wöhler. "Object classification in radar using ensemble methods." In: *2017 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*. IEEE. 2017, pp. 87–90.

[97] Martin Lowson. "Energy use and sustainability of transport systems." In: *Advanced Transport Group, University of Bristol, Tech. Rep* (2004).

[98] Wenjie Luo, Bin Yang, and Raquel Urtasun. "Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net." In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pp. 3569–3577.

[99] Mohamed Hussein Emam Mabrouk Mabrouk. "Signal Processing of UWB Radar Signals for Human Detection Behind Walls." PhD thesis. Université d'Ottawa/University of Ottawa, 2015.

[100] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. "1 year, 1000 km: The Oxford RobotCar dataset." In: *The International Journal of Robotics Research* 36.1 (2017), pp. 3–15.

[101] Ishrat Maherin and Qilian Liang. "A mutual information based approach for target detection through foliage using UWB radar." In: *2012 IEEE International Conference on Communications (ICC)*. IEEE. 2012, pp. 6406–6410.

[102] Julien Maitre, Kévin Bouchard, Camille Bertuglia, and Sébastien Gaboury. "Recognizing activities of daily living from UWB radars and deep learning." In: *Expert Systems with Applications* 164 (2021), p. 113994. ISSN: 0957-4174.

[103] Holger H Meinel and Wolfgang Bösch. "Radar sensors in cars." In: *Automated Driving*. Springer, 2017, pp. 245–261.

[104] J. M. Mendel. "Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications." In: *Proceedings of the IEEE* 79.3 (1991), pp. 278–305.

[105] Michael Meyer and Georg Kuschk. "Automotive radar dataset for deep learning based 3d object detection." In: *2019 16th European Radar Conference (EuRAD)*. IEEE. 2019, pp. 129–132.

[106] Amira Mimouna, Ihsen Alouani, Anouar Ben Khalifa, Yassin El Hillali, Abdelmalik Taleb-Ahmed, Atika Menhaj, Abdeldjalil Ouahabi, and Najoua Essoukri Ben Amara. "OLIMP: A Heterogeneous Multimodal Dataset for Advanced Environment Perception." In: *Electronics* 9.4 (2020), p. 560.

[107] Kazuki Minemura, Hengfui Liau, Abraham Monrroy, and Shinpei Kato. "LMNet: Real-time multiclass object detection on CPU using 3D LiDAR." In: *2018 3rd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*. IEEE. 2018, pp. 28–34.

[108] A. F. Molisch et al. "A Comprehensive Standardized Model for Ultrawideband Propagation Channels." In: *IEEE Transactions on Antennas and Propagation* 54.11 (2006), pp. 3151–3166.

[109] Basam Musleh, Arturo de la Escalera, and José María Armingol. "Uv disparity analysis in urban environments." In: *International Conference on Computer Aided Systems Theory*. Springer. 2011, pp. 426–432.

[110] Ramin Nabati and Hairong Qi. "RRPN: Radar Region Proposal Network for Object Detection in Autonomous Vehicles." In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 3093–3097.

[111] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. "Phased lstm: Accelerating recurrent network training for long or event-based sequences." In: *Advances in neural information processing systems* 29 (2016), pp. 3882–3890.

[112] Daniel Neumann, Tobias Langner, Fritz Ulbrich, Dorothee Spitta, and Daniel Goehring. "Online vehicle detection using Haar-like, LBP and HOG feature based image classifiers with stereo vision preselection." In: *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2017, pp. 773–778.

[113] Peishuang Ni, Chen Miao, Hui Tang, Mengjie Jiang, and Wen Wu. "Small Foreign Object Debris Detection for Millimeter-Wave Radar Based on Power Spectrum Features." In: *Sensors* 20.8 (2020), p. 2316.

[114] Felix Nobis, Maximilian Geisslinger, Markus Weber, Johannes Betz, and Markus Lienkamp. "A deep learning-based radar and camera sensor fusion architecture for object detection." In: *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE. 2019, pp. 1–7.

[115] Sang-Il Oh and Hang-Bong Kang. "Object detection and classification by decision-level fusion for intelligent vehicle systems." In: *Sensors* 17.1 (2017), p. 207.

[116] D. Orlando. "A Novel Noise Jamming Detection Algorithm for Radar Applications." In: *IEEE Signal Processing Letters* 24.2 (2017), pp. 206–210.

[117] Rafael Padilla, Sergio L Netto, and Eduardo AB da Silva. "A survey on performance metrics for object-detection algorithms." In: *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE. 2020, pp. 237–242.

[118] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." In: *International conference on machine learning*. 2013, pp. 1310–1318.

[119] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. "The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes." In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 9552–9557.

[120] Sujeet Milind Patole, Murat Torlak, Dan Wang, and Murtaza Ali. "Automotive radars: A review of signal processing techniques." In: *IEEE Signal Processing Magazine* 34.2 (2017), pp. 22–35.

[121] Margie Peden, Richard Scurfield, David Sleet, Adnan A Hyder, C Mathers, Eva Jarawan, AA Hyder, D Mohan, and E Jarawan. *World report on road traffic injury prevention*. World Health Organization, 2004.

[122] Maytus Piriyajitakonkij, Patchanon Warin, Payongkit Lakhan, Pitshaporn Leelaarporn, Nakorn Kumchaiseemak, Supasorn Suwajanakorn, Theerasarn Pianpanit, Nattee Niparnan, Subhas Chandra Mukhopadhyay, and Theerawit Wilaiprasitporn. "SleepPoseNet: Multi-view learning for sleep postural transition recognition using UWB." In: *IEEE Journal of Biomedical and Health Informatics* (2020).

[123] Alwin Poulose and Dong Seog Han. "UWB Indoor Localization Using Deep Learning LSTM Networks." In: *Applied Sciences* 10.18 (2020), p. 6290.

[124] Robert Prophet, Marcel Hoffmann, Martin Vossiek, Christian Sturm, Alicja Ossowska, Waqas Malik, and Urs Lübbert. "Pedestrian classification with a 79 GHz automotive radar sensor." In: *2018 19th International Radar Symposium (IRS)*. IEEE. 2018, pp. 1–6.

[125] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. "Frustum pointnets for 3d object detection from rgb-d data." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 918–927.

[126] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. "Pointnet: Deep learning on point sets for 3d classification and segmentation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[127] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." In: *Advances in neural information processing systems* 30 (2017), pp. 5099–5108.

[128] Zheng Qin, Zhaoning Zhang, Xiaotao Chen, Changjian Wang, and Yuxing Peng. "Fd-mobilenet: Improved mobilenet with a fast downsampling strategy." In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 1363–1367.

[129] Xuanjun Quan, Jeong Woo Choi, and Sung Ho Cho. "A New Thresholding Method for IR-UWB Radar-Based Detection Applications." In: *Sensors* 20.8 (2020), p. 2314.

[130] Rakesh Nattoji Rajaram, Eshed Ohn-Bar, and Mohan Manubhai Trivedi. "Looking at pedestrians at different scales: A multiresolution approach and evaluations." In: *IEEE Transactions on Intelligent Transportation Systems* 17.12 (2016), pp. 3565–3576.

[131] Thiago Rateke and Aldo von Wangenheim. "Road obstacles positional and dynamic features extraction combining object detection, stereo disparity maps and optical flow data." In: *Machine Vision and Applications* 31.7 (2020), pp. 1–11.

[132] 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*. 2016, pp. 779–788.

[133] Jimmy Ren, Xiaohao Chen, Jianbo Liu, Wenxiu Sun, Jiahao Pang, Qiong Yan, Yu-Wing Tai, and Li Xu. "Accurate single stage detector using recurrent rolling convolution." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5420–5428.

[134] 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*. 2015, pp. 91–99.

[135] Joon Hyo Rhee and Jiwon Seo. "Low-cost curb detection and localization system using multiple ultrasonic sensors." In: *Sensors* 19.6 (2019), p. 1389.

[136] Mark A Richards. *Fundamentals of radar signal processing*. McGraw-Hill Education, 2014.

[137] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. "Playing for data: Ground truth from computer games." In: *European conference on computer vision*. Springer. 2016, pp. 102–118.

[138] *Road causes*. https://www.nhtsa.gov/. 2020.

[139] *Road traffic injuries*. hhttps://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. 2020.

[140] Francisca Rosique, Pedro J Navarro, Carlos Fernández, and Antonio Padilla. "A systematic review of perception system and simulators for autonomous vehicles research." In: *Sensors* 19.3 (2019), p. 648.

[141] Rahmad Sadli, Charles Tatkeu, Khadija Hamidoun, Yassin El Hillali, and Atika Rivenq. "UWB radar recognition system based on HOS and SVMs." In: *IET Radar, Sonar & Navigation* 12.10 (2018), pp. 1137–1145.

[142] Hamidreza Sadreazami, Miodrag Bolic, and Sreeraman Rajan. "On the use of ultra wideband radar and stacked LSTM-RNN for at home fall detection." In: *2018 IEEE Life Sciences Conference (LSC)*. IEEE. 2018, pp. 255–258.

[143] L. Sakkila, P. Deloof, Y. Elhillali, A. Rivenq, and S. Niar. "A Real Time Signal Processing for an Anticollision Road Radar System." In: *IEEE Vehicular Technology Conference*. 2006, pp. 1–5. DOI: 10.1109/VTCF.2006.610.

[144] L Sakkila, Y Elhillali, A Rivenq, C Tatkeu, and JM Rouvaen. "Short range automotive radar based on UWB pseudo-random coding." In: *2007 7th International Conference on ITS Telecommunications*. IEEE. 2007, pp. 1–6.

[145] L Sakkila, Y Elhillali, J Zaidouni, A Rivenq, C Tatkeu, and JM Rouvaen. "High order statistic receiver applied to UWB radar." In: *2009 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*. IEEE. 2009, pp. 643–647.

[146] L. Sakkila, C. Tatkeu, F. Boukour, Y. El Hillali, A. Rivenq, and J. Rouvean. "UWB Radar system for road anti-collision application." In: *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*. 2008, pp. 1–6.

[147] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[148] Supriya B Sarkar and B Chandra Mohan. "Review on autonomous vehicle challenges." In: *First International Conference on Artificial Intelligence and Cognitive Computing*. Springer. 2019, pp. 593–603.

[149] Ole Schumann, Markus Hahn, Jürgen Dickmann, and Christian Wöhler. "Semantic segmentation on radar point clouds." In: *2018 21st International Conference on Information Fusion (FUSION)*. IEEE. 2018, pp. 2179–2186.

[150] Ole Schumann, Christian Wöhler, Markus Hahn, and Jürgen Dickmann. "Comparison of random forest and long short-term memory network performances in classification tasks using radar." In: *2017 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE. 2017, pp. 1–6.

[151] Romain Serizel, Victor Bisot, Slim Essid, and Gaël Richard. "Acoustic features for environmental sound analysis." In: *Computational analysis of sound scenes and events*. Springer, 2018, pp. 71–101.

[152] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. "Overfeat: Integrated recognition, localization and detection using convolutional networks." In: *arXiv preprint arXiv:1312.6229* (2013).

[153] Claude E Shannon. "A mathematical theory of communication." In: *Bell system technical journal* 27.3 (1948), pp. 379–423.

[154] W. Shao, T. R. McCollough, and W. J. Mccollough. "A Phase Shift and Sum Method for UWB Radar Imaging in Dispersive Media." In: *IEEE Transactions on Microwave Theory and Techniques* 67.5 (2019), pp. 2018–2027.

[155] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).

[156] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. "MVX-Net: Multimodal voxelnet for 3D object detection." In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 7276–7282.

[157] Akash Singh. *Anomaly detection for temporal data using long short-term memory (lstm)*. 2017.

[158] Shuran Song and Jianxiong Xiao. "Deep sliding shapes for amodal 3d object detection in rgb-d images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 808–816.

[159] 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*. 2015, pp. 1–9.

[160] Yongting Tao and Jun Zhou. "Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking." In: *Computers and electronics in agriculture* 142 (2017), pp. 388–396.

[161] *The Effidence Organization*. https://www.effidence.com/. 2018.

[162] Jitendra K Tugnait. "Time delay estimation with unknown spatially correlated Gaussian noise." In: *IEEE Transactions on Signal Processing* 41.2 (1993), pp. 549–558.

[163] *UMAIN*. http://www.umain.co.kr/en/uwb-radar-sensor-module/. 2020.

[164] *Udacity Self-Driving Car*. https://github.com/udacity/self-driving-car. 2019.

[165] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. "Selective search for object recognition." In: *International journal of computer vision* 104.2 (2013), pp. 154–171.

[166] Victor Vaquero, Ely Repiso, Alberto Sanfeliu, John Vissers, and Maurice Kwakkernaat. "Low cost, robust and real time system for detecting and tracking moving objects to automate cargo handling in port terminals." In: *Robot 2015: Second Iberian Robotics Conference*. Springer. 2016, pp. 491–502.

[167] Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features." In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. IEEE. 2001, pp. I–I.

[168] Theo Vos, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, et al. "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019." In: *The Lancet* 396.10258 (2020), pp. 1204–1222.

[169] Jörg Wagner, Volker Fischer, S. Herman, and Sven Behnke. "Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks." In: *ESANN*. 2016.

[170] Safat B Wali, Mahammad A Hannan, Aini Hussain, and Salina A Samad. "An automatic traffic sign detection and recognition system based on colour segmentation, shape matching, and svm." In: *Mathematical Problems in Engineering* 2015 (2015).

[171] Jian-Gang Wang, Simon Jian Chen, Lu-Bing Zhou, Kong-Wah Wan, and Wei-Yun Yau. "Vehicle detection and width estimation in rain by fusing radar and vision." In: *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE. 2018, pp. 1063–1068.

[172] Xiao Wang, Linhai Xu, Hongbin Sun, Jingmin Xin, and Nanning Zheng. "On-road vehicle detection and tracking using MMW radar and monovision fusion." In: *IEEE Transactions on Intelligent Transportation Systems* 17.7 (2016), pp. 2075–2084.

[173] Jian Wei, Jianhua He, Yi Zhou, Kai Chen, Zuoyin Tang, and Zhiliang Xiong. "Enhanced object detection with deep convolutional neural networks for advanced driving assistance." In: *IEEE Transactions on Intelligent Transportation Systems* 21.4 (2019), pp. 1572–1583.

[174] Xinshuo Weng and Kris Kitani. "Monocular 3d object detection with pseudo-lidar point cloud." In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019, pp. 0–0.

[175] Paul J Werbos. "Backpropagation through time: what it does and how to do it." In: *Proceedings of the IEEE* 78.10 (1990), pp. 1550–1560.

[176] Sascha Wirges, Tom Fischer, Christoph Stiller, and Jesus Balado Frias. "Object detection and classification in occupancy grid maps using deep convolutional networks." In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2018, pp. 3530–3535.

[177] Magnus Wrenninge and Jonas Unger. "Synscapes: A photorealistic synthetic dataset for street scene parsing." In: *arXiv preprint arXiv:1810.08705* (2018).

[178] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. "Subcategory-aware convolutional neural networks for object proposals and detection." In: *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2017, pp. 924–933.

[179] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. "Pointfusion: Deep sensor fusion for 3d bounding box estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 244–253.

[180] Jianru Xue, Jianwu Fang, Tao Li, Bohua Zhang, Pu Zhang, Zhen Ye, and Jian Dou. "BLVD: Building a large-scale 5d semantics benchmark for autonomous driving." In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 6685–6691.

[181] Yan Yan, Yuxing Mao, and Bo Li. "Second: Sparsely embedded convolutional detection." In: *Sensors* 18.10 (2018), p. 3337.

[182] Degui Yang, Zhengliang Zhu, and Buge Liang. "Vital sign signal extraction method based on permutation entropy and EEMD algorithm for ultra-wideband radar." In: *IEEE Access* 7 (2019), pp. 178879–178890.

[183] Fan Yang, Wongun Choi, and Yuanqing Lin. "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2129–2137.

[184] Lina Yang, Yingping Huang, Xing Hu, Hongjian Wei, and Qixiang Wang. "Multiclass obstacles detection and classification using stereovision and Bayesian network for intelligent vehicles." In: *International Journal of Advanced Robotic Systems* 17.4 (2020), p. 1729881420947270.

[185] J Javier Yebes, Luis M Bergasa, and Miguel García-Garrido. "Visual object recognition with 3D-aware features in KITTI urban scenes." In: *Sensors* 15.4 (2015), pp. 9228–9250.

[186] Zhendong Yin, Kai Cui, Zhilu Wu, and Liang Yin. "Entropy-based TOA estimation and SVM-based ranging error mitigation in UWB ranging systems." In: *Sensors* 15.5 (2015), pp. 11701–11724.

[187] S. Yoo, S. Chung, D. Seol, and S. H. Cho. "Adaptive Clutter Suppression Algorithm for Detection and Positioning using IR-UWB Radar." In: *2018 9th International Conference on Ultrawideband and Ultrashort Impulse Signals (UWBUSIS)*. 2018, pp. 40–43.

[188] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. "Bdd100k: A diverse driving video database with scalable annotation tooling." In: *arXiv preprint arXiv:1805.04687* 2.5 (2018), p. 6.

[189] Xue Yuan, Shuai Su, and Houjin Chen. "A graph-based vehicle proposal location and detection algorithm." In: *IEEE Transactions on Intelligent Transportation Systems* 18.12 (2017), pp. 3282–3289.

[190] Jianming Zhang, Zhipeng Xie, Juan Sun, Xin Zou, and Jin Wang. "A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection." In: *IEEE Access* 8 (2020), pp. 29742–29754.

[191] Yin Zhou and Oncel Tuzel. "Voxelnet: End-to-end learning for point cloud based 3d object detection." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4490–4499.

[192] Adam Ziebinski, Rafal Cupek, Hueseyin Erdogan, and Sonja Waechter. "A survey of ADAS technologies for the future perspective of sensor fusion." In: *International Conference on Computational Collective Intelligence*. Springer. 2016, pp. 135–146.

[193] SAE international. “Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles.” In: *SAE International,(J3016)* (2016).