

CRANFIELD UNIVERSITY

**Visual/Acoustic Detection and
Localisation in Embedded Systems**

Riad AZZAM

Supervisor:

Dr.Nabil AOUF

PhD

August 2015

Abstract

The continuous miniaturisation of sensing and processing technologies is increasingly offering a variety of embedded platforms, enabling the accomplishment of a broad range of tasks using such systems. Motivated by these advances, this thesis investigates embedded detection and localisation solutions using vision and acoustic sensors. Focus is particularly placed on surveillance applications using sensor networks. Existing vision-based detection solutions for embedded systems suffer from the sensitivity to environmental conditions. In the literature, there seems to be no algorithm able to simultaneously tackle all the challenges inherent to real-world videos.

Regarding the acoustic modality, many research works have investigated acoustic source localisation solutions in distributed sensor networks. Nevertheless, it is still a challenging task to develop an efficient algorithm that deals with the experimental issues, to approach the performance required by these systems and to perform the data processing in a distributed and robust manner. The movement of scene objects is generally accompanied with sound emissions with features that vary from an environment to another. Therefore, considering the combination of the visual and acoustic modalities would offer a significant opportunity for improving the detection and/or localisation using the described platforms.

In the light of the described framework, we investigate in the first part of the thesis the use of a cost-effective visual based method that can deal robustly with the issue of motion detection in static, dynamic and moving background conditions. For motion detection in static and dynamic backgrounds, we present the development and the performance analysis of a spatio-temporal form of the Gaussian mixture model. On the other hand, the problem of motion detection in moving

backgrounds is addressed by accounting for registration errors in the captured images. By adopting a robust optimisation technique that takes into account the uncertainty about the visual measurements, we show that high detection accuracy can be achieved.

In the second part of this thesis, we investigate solutions to the problem of acoustic source localisation using a trust region based optimisation technique. The proposed method shows an overall higher accuracy and convergence improvement compared to a linear-search based method. More importantly, we show that through characterising the errors in measurements, which is a common problem for such platforms, higher accuracy in the localisation can be attained.

The last part of this work studies the different possibilities of combining visual and acoustic information in a distributed sensors network. In this context, we first propose to include the acoustic information in the visual model. The obtained new augmented model provides promising improvements in the detection and localisation processes. The second investigated solution consists in the fusion of the measurements coming from the different sensors. An evaluation of the accuracy of localisation and tracking using a centralised/decentralised architecture is conducted in various scenarios and experimental conditions. Results have shown the capability of this fusion approach to yield higher accuracy in the localisation and tracking of an active acoustic source than by using a single type of data.

Acknowledgements

I Am grateful to the following for contributing to this thesis in one way or another. To Dr. Nabil Aouf for his guidance, motivation and technical advice. To my committee members, especially Professor Mark Richardson for his advice.

To Cranfield Univeristy and the Shrivenham Defence Academy for providing an environment conducive to research.

A special thanks to Tarek, Saif, Lounis, Oualid for their support through difficult times in research and otherwise. Without you this experience would have been less fulfilling. A very special thanks to Tarek and Oualid for proof reading the first draft of this thesis.

To my colleagues Mohammed, Badis, Rahim, Ahmed, Abdennour.

To friends and relatives who have kept me in their prayers. To my parents, sisters and brothers who, although being so far away, have been with me every step of the way.

Many thanks to my wife for your enduring patience and being so helpful and supportive, and to my little kids, you are my everyday motivation.

Contents

Abstract	ii
Acknowledgements	iv
List of Figures	x
List of Tables	xiv
1 Introduction	1
1.1 Research motivation	4
1.2 Thesis organisation and contribution	5
1.3 Contribution	7
1.4 Software Tools	8
2 Image representation, visual detection and acoustic source localisation	11
2.1 Introduction	11
2.2 Video detection	12
2.2.1 Image representation	12
2.2.2 Embedded systems and Data compression techniques	14
2.2.3 Pixel based motion detection	15
2.2.4 Feature based motion detection	16
2.2.5 Background segmentation post-processing	16
2.2.6 Connected-component labelling	17
2.2.7 Fundamental concepts for image projection	18
2.3 Acoustic source localisation	21
2.3.1 Physics of sound	21
2.3.2 Amplitude, frequency, wavelength and velocity	22
2.3.3 Sound field	24
2.3.4 Basic Acoustics Features	25
2.3.5 Features in the time domain	25

2.3.6	Features in the frequency domain	26
2.3.7	Methods to analyse audio signals	27
2.3.7.1	Fourier Transform	27
2.3.7.2	Time-frequency analysis	28
2.3.7.3	The continuous wavelet transform	28
2.3.8	The generalised cross correlation	29
2.3.9	Acoustic source localisation in wireless sensor networks	29
2.3.10	Architectures of distributed systems for acoustic localisation	31
2.3.10.1	Decentralised systems	31
2.3.10.2	Centralised systems	31
2.3.11	Middle-Ware services	32
2.3.12	The Signal Model	33
2.3.12.1	The energy based approach	33
2.3.12.2	Time delay of arrival (TDOA)	34
3	Visual Object Detection with Spatially Global Gaussian Mixture	36
3.1	Introduction	36
3.2	Related Works	38
3.2.1	Gaussian mixture models	38
3.2.2	Adaptive background learning models	40
3.2.3	Advanced statistical models	41
3.2.4	Low-rank minimisation models	42
3.3	Description of the spatially global gaussian mixture model	43
3.3.1	Background model estimation	44
3.3.2	Background model update	47
3.4	SGGMM Based Pixel Location Uncertainties Approach	49
3.4.1	Computation of pixel displacement uncertainties	50
3.4.2	Statistics computation of pixel displacement uncertainty	53
3.4.3	Using pixel uncertainty in background subtraction	55
3.5	Foreground segmentation	56
3.6	Performance Tests	56
3.6.1	Performance evaluation of the SGGMM model	56
3.6.2	Quantitative evaluation	57
3.6.3	Qualitative evaluation	61
3.6.3.1	Segmentation using colour based SGGMM	61
3.6.4	Segmentation using SGGMM with colour and pixel uncertainties	64
3.6.5	Computation performance and real time sensor node implementation	67
3.7	Conclusion	74

4	Moving Object Detection from a Moving platform	76
4.1	Introduction	76
4.2	Related work	77
4.3	Structure of the proposed solution	79
4.4	Robust homography matrix estimation	81
4.4.1	Feature detection with uncertainty	81
4.4.2	The covariance intersection (CI)	83
4.4.3	Modelling the problem of the hHomography matrix estimation	83
4.4.4	The H_∞ filter	85
4.5	Motion detection using Optical Flow	89
4.5.1	Estimation of the optical flow using local methods	89
4.5.2	Modelling optical flow using the SGGMM	93
4.6	Experimental results	95
4.7	Conclusion	100
5	Robust Acoustic Source Localisation in Low Cost Sensor Networks	102
5.1	Introduction	102
5.2	Related works	103
5.3	TDOA based localisation signal model In WSN	105
5.4	Uncertainty in TDOA measurements in Low cost WSN	108
5.4.1	The Uncertainties due to synchronisation issue	108
5.4.2	The uncertainty due the low sampling frequency	111
5.4.3	Estimation of the drift rate between notes	112
5.5	Powells Dogleg/Double Dogleg optimisers	115
5.5.1	Powell's method with Double Dogleg step	120
5.5.2	Weighting norms adoption	122
5.5.3	The total least squares (TLS)	123
5.6	Experimental setup and experiments	124
5.6.1	Performance evaluation of the the proposed method	126
5.6.1.1	Convergence rate and execution time	128
5.7	Conclusion	133
6	Active Acoustic Sources Localisation in Distributed Sensor Networks	135
6.1	Introduction	135
6.2	Related works	136
6.3	First part: Augmented SGGMM with the acoustic information	138
6.3.1	Proposed Fusion architecture	138
6.3.2	Time of acoustic events measurements using the Micaz sensor notes	142

6.3.3	Acoustic source localisation with outliers and erroneous measurements elimination	143
6.3.4	Dealing with time delay of measurements arrival	145
6.3.5	Including the acoustic information in the SGGMM model	146
6.3.6	Experimental tests	147
6.3.6.1	Evaluation of the improvement in the detection accuracy	147
6.3.6.2	Evaluation of the improvement in localisation accuracy	150
6.4	Second part: Cooperative localisation and tracking	154
6.4.1	Fusion based on centralised/decentralised architecture	155
6.4.2	Fusion at the central level	157
6.4.3	The motion model	158
6.4.4	Local level of tracking	159
6.4.4.1	Local level tracking using the acoustic data	159
6.4.4.2	Problem modelling	160
6.4.4.3	The Unscented transformation (UT)	161
6.4.4.4	The UKF Algorithm	162
6.4.4.5	Results and discussions	164
6.4.4.6	Local level of tracking using the visual Data	165
6.4.5	Experimental setup	166
6.4.6	Experiments	168
6.4.7	Evaluation result	170
6.4.7.1	For a stationary acoustic source	170
6.4.7.2	Moving in a linear path	171
6.4.7.3	Motion with constant turn rate	172
6.5	Conclusion	174
7	Conclusion and Future Work	176
	Bibliography	180

List of Figures

2.1	Illustration of an image matrix representation	13
2.2	Illustration of image colour representations	13
2.3	Camera world reference transformation	19
2.4	Representation of a sound wave ¹	23
2.5	Wavelength versus frequency under normal conditions(air) ²	24
2.6	Sensor node architecture	30
3.1	Different components of the SGMM based model	50
3.2	Effect of the dynamic background on accurate detection	51
3.3	Background subtraction approach based on pixel uncertainties	51
3.4	Number of test frames for the evaluation of the SGMM	58
3.5	Pixel location uncertainties normalised histogram of the background points: (a) pixel (100,100) of the first scene, (b) pixel (100,100) of the secone scene, (c) pixel (100,100) of the third scene, (d) pixel (100,100) of the fourth scene	59
3.6	Comparison between binary objects mask of each method for Office sequence	62
3.7	Comparison between binary masks of each method for the PETS2006 sequence	63
3.8	Comparison between binary masks of each method for Canoe sequence	65
3.9	Comparison between binary mask of each method for the Overpass sequence	66
3.10	Execution times of the studied algorithms	67
3.11	The embedded camera used in the experiments with its major components	68
3.12	Software architecture handling the CITRIC camera board	69
3.13	Number of test frames for the evaluation of the SGMM using the CITRIC camera	70
3.14	Comparison between binary masks of the SGMM based colour only and SGMM (W.P.U)	71
3.15	Evaluation results of the SGMM based model using for UASL sequence	72

3.16	Evaluation results of the SGGMM based model using for Garden sequence	72
3.17	Evaluation results of the SGGMM based model using for Street sequence	73
4.1	The overall architecture of the proposed solution	80
4.2	Matched features and their corresponding Uncertainties from RGB channels ³	82
4.3	Estimating the Homography matrix using the robust H_∞	89
4.4	Optical flow estimation for dynamic background: a) first image, b) second image, c) the u element of the velocity vector, d) the v element of the velocity vector	92
4.5	Proposed scheme for moving object detection	94
4.6	Qualitative evaluation of the proposed method-first scenario, first test ⁴	96
4.7	Qualitative evaluation of the proposed method-first scenario, second test	97
4.8	Qualitative evaluation of the proposed method-second scenario, first test	98
4.9	Qualitative evaluation of the proposed method-second scenario, second test	99
6.1	Different components of the fusion architecture proposed	139
6.2	Software architecture of the proposed solution implemented at the CITRIC camera	141
6.3	Architecture of the Nesc programme in charge of acoustic event detection	143
6.4	Propagation of localisation errors in a distributed acoustic sensors network composed of 7 nodes with erroneous measurements.	144
6.5	First test scenario with the corresponding ground truth.	147
6.6	Second test scenario with the corresponding ground truth.	148
6.7	Third test scenario with the corresponding ground truth.	148
6.8	Results of SGGMM colour background model and augmented SGGMM with the acoustic signal for active acoustic object detection.	149
6.9	Quantitative evaluation of the improvement in detection	150
6.10	Setup used for the robot localisation using the proposed fusion approach	151
6.11	Moving platform trajectory with the collected measurements	152
6.12	Accuracy Comparison between localisation based on acoustic measurements only and SGGMM based acoustics	153
6.13	Main component of the centralised/decentralised architecture of fusion.	156
6.14	Illustration of the unscented transform(UT)	162

6.15	The alternation between the UKF steps	163
6.16	Behaviour of the UKF in reaching the optimum solution for the sound source localisation	164
6.17	UKF feeding with visual measurements.	165
6.18	Setup used for the centralised/decentralised architecture of fusion implemented	167
6.19	Results obtained from the tracking scheme applied to the different motion models: a:Stationary, b:Linear motion, c:Circular trajectory	169
6.20	MSE obtained using different algorithm for static object position estimation	171
6.21	MSE obtained using different algorithm for tracking the active acoustic object following linear trajectory	172
6.22	MSE recorded using the different algorithms for tracking the active acoustic object following circular trajectory	173

.

List of Tables

2.1	Image and Video Compression Standards	14
3.1	Comparison between the Average Similarity, F-measure, Precision, and Recall values for each method	60
4.1	Average back projection error estimation obtained from using dif- ferent methods	88
4.2	Qualitative evaluation of the proposed scheme	95

Chapter 1

Introduction

Surveillance and monitoring of public and private spaces is progressively becoming a very important and critical issue, particularly after the recent burst of terrorist attacks. It is therefore imperative that effective surveillance systems are developed to ensure high security levels. Ideally, different sensors are employed to accomplish this mission. In this context, micro-devices technology witnessed a significant development which yielded the appearance of the micro-sensing and actuation devices. Such advances have revolutionised the way engineers understand and manage complex physical systems. Notably, the capabilities of detailed physical monitoring and manipulation offered enormous opportunities, not only for surveillance systems, but for most scientific disciplines by providing embedded processing platforms with exciting capabilities.

Using such technology allows us to carry out surveillance missions in unfriendly environments such as remote geographic regions or toxic locations. Furthermore, it enables sensing and maintenance in large industrial plants, military surveillance and combat operations. In practically most of these applications, key requirements include robustness with regard to different disturbances and uncertainties, adaptation to different environments, as well as optimal consumption of resources to ensure permanent operability. In general, the proposed surveillance systems

integrate different modalities of sensors to ensure higher accuracy and/or permanent operability. These systems can be organised in varieties of architectures with the sensors placed at the first level. Detection is the primary operation of every surveillance activity.

Detected changes by each physical sensor which processed to extract meaningful features, and then may be integrated to recover missing data. This integration enables increased improvement at higher levels of processing. It involves either increasing the accuracy of object localisation, which is based on the ground plane hypothesis or object recognition. Such processing may include the tracking of each detected feature in the scene on the image plane and transform the 2D blob positions (in the sensor coordinates frame) into 3D object positions (in the coordinates of the monitored environment's map) using the visual sensor. It may also include the inverse operation when using the acoustic sensors.

Reliance on wireless connectivity is crucial for surveillance activity since for most envisaged applications, the observed environment does not have adequate infrastructure for either communication or energy supply. Hence, untethered nodes must rely on small local power sources and wireless communication channels. The design of new generations of smart video camera for motion detection enabled in-network processing of images to reduce the communicational load which has traditionally been high in existing camera networks with centralised processing. These camera nodes enable a broad range of distributed applications compared to traditional platforms. Indeed they provide more computing capabilities and tighter integration of physical components while still consuming relatively low power. In the same context of video surveillance, improvement in efficiency has greatly improved by the introduction of the unmanned aerial platforms to complete a monitoring mission within a designated territory or early threat detection for local security.

The use of distributed sensor networks (DSN) for the location of acoustic sources

has been of a great benefit to different areas such as intruder detection, sniper localisation, automatic tracking of acoustic source or speakers in an e-conferencing environment and to voice enhancement. The idea behind acoustic source localisation systems is the use of multiple sensors (microphone arrays) placed at different known location. Since sound travels with a constant speed from the sound source to the sensors, the recorded signals can be used to estimate the possible location of the source.

The integration of various modalities in surveillance systems is carried out using data fusion techniques. The aim is to increase both the range of detection and then capability to detect interesting events. Multisensory surveillance systems can take advantage of either same type of information acquired from different spatial locations or information acquired by sensors of different types. Appropriate processing techniques and new sensors providing real-time information related to different scene characteristics can to enlarge the size of monitored environments and to improve performances in terms of activity detection over the monitored areas.

Another objective of data fusion is to ensure accurate target tracking, an activity which has increased in popularity in distributed sensor networks (DSN). This is mainly due the reduced cost of sensors, which led to: the possibility of deploying large number to achieve wide area coverage, and to increase the density allowing sensors to reside far closer to the objects being sensed. Additionally, improvement in sensing quality with overlapping coverage resulted in increased robustness and improved accuracy. Furthermore, the diversity of sensing modalities offers new solutions based on complementary configurations; for instance, certain types of sensors (e.g., microphone arrays, radars) provide good ranging data, while others (e.g., cameras, Infrared(IR)) are ideal for object orientation and classification. This diversity in sensing modalities is exploited to provide accurate and rich information about the target. It is also important to note that the spatial sensing diversity greatly mitigates the effects of obstructions on line-of-sight sensors.

1.1 Research motivation

The main motivation behind this work is the development and implementation of detection and localisation algorithms. The targets of interest can be either moving objects, stationary or moving sound sources. More specifically, we investigate the problem of accurate detection and localisation in embedded systems by studying the visual and acoustic modalities.

In computer vision, solutions based on change detection represent a fundamental pre-processing step for embedded vision detection. However, most of the existing solutions are sensitive to environmental conditions such as illumination variations. The algorithms available in the literature seem to be unable to simultaneously address all the key challenges that undermine real-world videos. Moreover, there is a lack of realistic large-scale datasets, which cover real-world challenges and include accurate ground truths.

One of the most challenging tasks in acoustics lies in the determination of the sound source especially when a significant portion of the measurements is corrupted with noise from unknown sources. Since the early nineties, a number of standard and highly functional methods based on microphone arrays have matured. Today, with the newly introduced distributed sensor networks, which feature services that are executed within tightly constrained conditions, new solutions need to be investigated. The work we propose in part of this thesis aims to achieve this goal while taking into account the embedded constraints in acoustic sensor nodes.

Multi-sensory systems have shown their ability in improving the overall performance compared to mono-sensory configurations. In particular, the possibility of registering acoustic and vision measurements in a common coordinate system enables achieving an important improvement. Therefore, data fusion strategies

targeted by this research would offer a promising opportunity to ameliorate the detection and localisation accuracy in distributed platforms.

The described situations motivated the investigation of solutions based on both data modalities (visual and acoustic) where hardware implementation was a primary concern. The accuracy in the detection and localisation constitutes an important objective of this thesis.

1.2 Thesis organisation and contribution

This thesis is concerned with the investigation and development of efficient and accurate detection and localisation solutions for embedded systems in both outdoor and indoors environments. The principal contributions are made towards reliable data exploration, data filtering and fusion methodologies across platforms. A brief summary of the contributions presented in this thesis is as follows:

Theoretical background: On Chapter 2, we present the background required to carry out this research. We introduce the fundamental concepts in vision, vision detection process, fundamental concepts in acoustics. In addition, we discuss the corresponding approach for acoustic localisation in the DSNs.

Visual detection in static camera: On chapter 3, we investigate the problem of visual detection of moving objects using a model based on the Gaussian mixture models (GMM). The presented method, the Spatio-temporal Global Gaussian Mixture Model (SGGMM) uses RGB and Pixel Uncertainty for background modelling. The SGGMM (with colours only) is used for scene with moderate illumination changes. By including the pixel uncertainty statistics in the background model, the method can deal efficiently with dynamic backgrounds and backgrounds with fast luminosity variations.

We also handle experimental evaluation in indoor and outdoor environments which show the performance of foreground segmentation (object detection) with the proposed SGGMM model. These experimental scenarios take into account changes in the background within the scene. They are also used to compare the proposed technique with other state-of-the-art segmentation approaches in terms of accuracy and executions performance. To further confirm the latter, our solution is implemented and tested on an embedded camera.

Visual detection from moving camera: On Chapter 4, we investigate the problem of motion detection from a moving camera system. This is motivated by of the wide range of applications for moving object detection using moving platforms. These vary from security enhancement for borders and public spaces (using aerial platforms and PTZ cameras), to applications for industrial and daily activities, such as used for mobile robots and vehicles and driver assistance. Unlike motion detection using static cameras, and despite the considerable efforts made to investigate this problem, only few proposals are reported in the literature to efficiently address the challenging task of detecting in such scenarios. In this work we present an approach based on affine image warping using a robust method of homography for motion compensation and optical flow.

Robust acoustic source localisation in WSN: On Chapter 5, we aim to develop an efficient and robust algorithm for acoustic source localisation based on the Time Delay Of Arrival (TDOA) measurements. This algorithm is to be used in a context of low-cost sensor networks. Part of the available solutions in the literature formulate this problem as a minimisation of a non-linear least square function, which is solved using Gauss-Newton method. The latter shows a degraded performance especially when it is initialised far away from the desired solution. To make up for this inefficiency, we propose to adapt a trust region based optimiser named Double Dogleg.

Furthermore, we characterise, the uncertainties available in the TDOA measurements and propose a new way of evaluating them experimentally. These uncertainties are taken formally into account in the proposed optimiser through the adoption of weighted norms in its optimisation process. Evaluation results based on a source localisation setup demonstrate the suitability of the proposed algorithm in terms of the overall accuracy and the global convergence rate.

Detection, localisation and tracking in heterogeneous distributed networks: Chapter 6 is devoted to the problem of improved detection and localisation in heterogeneous distributed sensor networks. Relying on the correlation between the two data modalities (acoustic and video), we firstly propose an innovative solution for active acoustic source detection and localisation for a distributed sensor networks. This solution aims to augment the RGB vector used in the Spatio-temporal GGMM background subtraction method proposed in Chapter 3. This augmentation is done with the use of acoustic information to detect possible moving sound sources. A second contribution in this chapter concerns investigating the design and performance of the fusion based on the centralised/centralised architectures. The aim for such fusion approach is to evaluate the quality of tracking of active sound sources with regards to the communicational cost and the fusion algorithm used in the distributed sensor networks.

1.3 Contribution

- R.Azzam and N.Aouf "Acoustic detection and localisation enhanced by video analysis". Published by IEEE SMC (System, Man, Cybernetics). Manchester, October 2012.
- R.Azzam and N.Aouf "The Gaussian Processes for Acoustic Localisation and Tracking in Wireless Sensor Network", published at IET ICDP, imaging for crime detection and prevention , London, December 2013.

- R.Azzam and N.Aouf, "Acoustics And Video Fusion In Wireless sensors networks" in Cranfield Symposia-Shrivenham-Campus, June 2014.
- R.Azzam and N.Aouf "a Non-Parametric Tool for Vision Detection Analysis", published at IEEE Electronic Martime conference, September 2014.
- R.Azzam and N.Aouf "Embeded Fusion of Visual and Acoustic for Active Acoustic Source Detection With SGGMM" published at IEEE Electronic Martime conference, September 2014.
- R. Azzam, N.Aouf, M.Kemmouche, and M.Richardson "Efficient Visual Object Detection with Spatially Global Gaussian Mixture Models and Uncertainties" Under review after second round of revision, Journal of visual communication and image representation.
- R.Azzam and N.Aouf "Robust Non-linear Squares Optimiser for acoustic source localisation in WSN", Under review in Journal of Applied Acoustics.
- R.Azzam and N.Aouf "Optical Flow Based GMM and Robust Homography for Moving Object Detection in Moving Background", under review in journal of signal image and video processing.

1.4 Software Tools

Listed are the tools used during the study:

- Matlab: A technical computing environment developed by the MathWorks company;
- C/C++: General-purpose programming language, comprises both high-level and low-level language features.

-
- Nesc (Network Embedded systems C): Is a component-based, event-driven programming language used to build wireless sensor networks applications.
 - TinyOs: Is an open source, BSD (Berkeley Software Distribution) licensed operating system designed for low-power wireless devices, such as those used in sensor networks.
 - OpenCV: An open source library of programming functions aimed at real-time computer vision applications developed by Intel and supported by Willow Garage.

This page is intendedly left blank

Chapter 2

Image representation, visual detection and acoustic source localisation

2.1 Introduction

This Chapter is devoted to introducing basic concepts of visual and acoustic detection that serve as a foundation for the main contributions of this thesis. It involves the basic methods used for detection using static cameras, blob analysis methods, in addition to the so-called pinhole camera model [1, 2]. Using this model, the positions of the detected moving objects can be estimated for further processing. In this Chapter, we also include the basic concepts used in acoustics signal processing and the different approaches for acoustic localisation in distributed sensors networks.

2.2 Video detection

In automatic video surveillance systems, visual detection and localisation is the process of using visible information to detect and localise moving targets. Target classification is also among the most debated applications in this field. Performing target detection and tracking using a video stream leads to taking advantage of the pixel value or features, therefore a description of the image representation will be given first in the following section.

2.2.1 Image representation

An image is a multidimensional signal acquired from light captured by using digital camera. It is defined as a matrix where each entry (named pixel) is defined by a 2D index (i,j) ; i for the columns and j for the rows (see Figure 2.1). A pixel stores a value representing the corresponding intensity of the acquired light. The intensity of a pixel takes a value from 0 to 255. Image capability is characterised by its resolution, which latter has the following two aspects:

Spatial resolution representing the number of pixels (or matrix elements) of the image covering the visual space of the capture images. It corresponds to the product of the image columns and rows. The size of the spatial resolution has a quality effect on the projection of the captured scene into an image. The larger the resolution, the better is the image quality. However, a higher resolution implies bigger size which may lead to demand in storage capacity. A colour image corresponds to a 24 bits ($3 \times 8bits$) where each channel represents a primary colour namely red, green, and blue (RGB). Figure 2.2 gives a presentation of these two colour representations.

Temporal resolution: corresponding the number of images continuously acquired in a given time for a video device for instance. It is expressed as frames

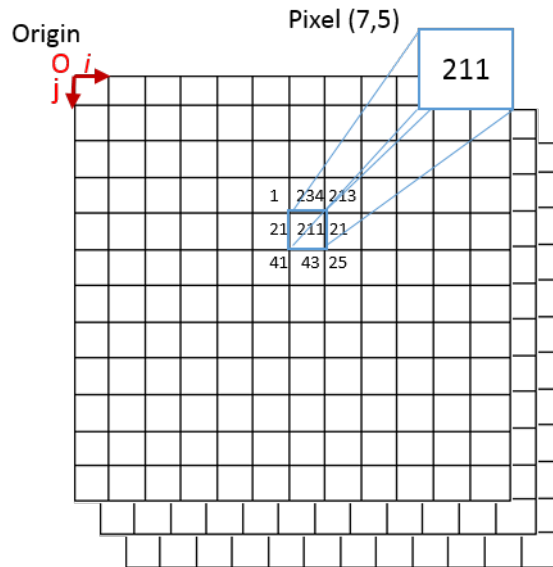


FIGURE 2.1: Illustration of an image matrix representation

per second (fps). The experiments and tests carried out in this thesis are using mainly RGB images.

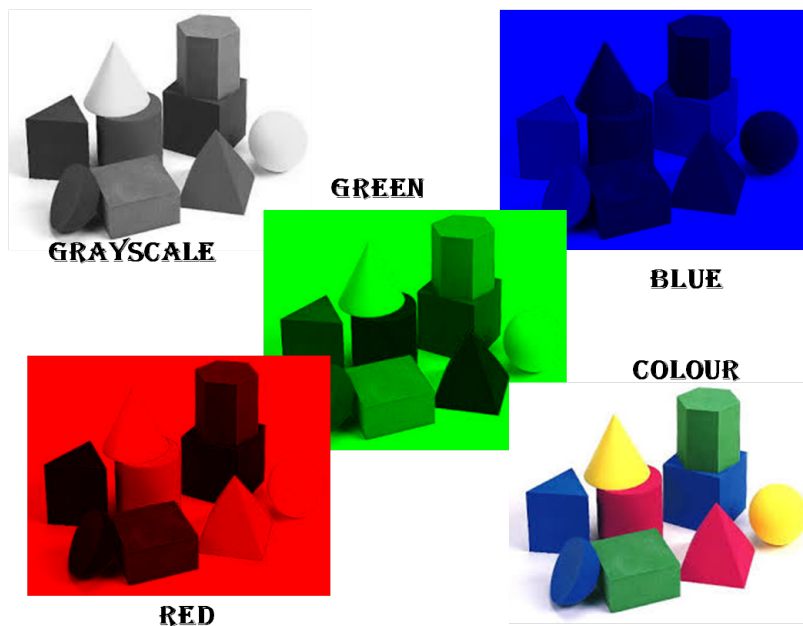


FIGURE 2.2: Illustration of image colour representations

2.2.2 Embedded systems and Data compression techniques

Unlike the general-purpose computer, which is devoted to manage a wide range of processing tasks, embedded system is dedicated computer system designed for specific and limited (generally one or two) functions. Their system is embedded as a part of a complete device system which includes hardware, such as electrical and mechanical components [3]. Embedded systems are resource constrained i.e., they generally have limited memory and computational capabilities and there is a driving need to extract as much space efficiency and performance from the available resources as possible. Code compression addresses both of these requirements [4]. Compressing the application binary and decompressing it at runtime enables better utilisation of the limited memory space in embedded systems. To reduce the costs associated with the large data size, three commonly used methods of compression are well known and reported in literature, these are [4]: Compiler-based [5], Instruction set compaction [6] and (lossless or lossy) data compression techniques [7].

Standard	Application	Bit rate
JPEG	Still image compression, CCTV	Variable
H.261	Video conference over ISDN	Multiple of 64 Kb/s
MPEG-1	Video on digital storage media (CD-ROM).	1.5 Mb/s
MPEG-2	Digital Television, CCTV.	2-20 Mb/s
H.263	Video telephony over PSTN.	≥ 33.6 Kb/s
MPEG-4	Object-Based Coding, synthetic content, Interactivity. CCTV.	Variable
JPEG-200	Improved still image compression.	Variable
H.264/ MPEG-4 AVC	Improved video compression.	10's to 100's Kb/s
Wavelet	CCTV recording	20–256 Kbits/s

TABLE 2.1: Image and Video Compression Standards

The importance of such techniques in distributed video (surveillance) system, is trivial, since it enables reducing the data rates for storing and transmitting video sequences. Different standard has been used in video compression techniques. The most known until nowadays are given Table 2.1 [8]. Each of them is suited for specific applications.

2.2.3 Pixel based motion detection

In surveillance applications, video detection and localisation is the process of using visible data for the detection and localisation of moving targets. Performing target detection and tracking using a video stream requires taking advantage of pixels values. The basic principle is based on detecting the foreground objects as the difference between the a pixel value in the current frame ($Frame_i$) and its corresponding in an image of the scene's static background ($Background_i$):

$$|Frame_i - Background_i| > Th \quad (2.1)$$

This principle is used to obtain the image of the scene's static Background automatically, while taking into account other parameters such as the change of the illumination and, natural motion of objects belonging to the background (motion of tree branches and leaves in the background). Many techniques have been proposed in the literature to tackle this problem, from which the mixture of Gaussians (or the Gaussian mixture models). The latter has gained much attention due to its low computation resources requirement, with reasonable results in real-time applications [9]. A detailed literature review on background foreground methods will be provided in the next Chapter.

2.2.4 Feature based motion detection

This encloses the class of methods based on the optical flow [10–13]. The latter is the distribution of apparent velocities of movement of brightness patterns in an image. Optical flow can arise from relative motion of objects with respect to the viewer. Consequently, optical flow can give important information about the spatial arrangement of the imaged objects and the rate of change of this arrangement [10]. Different techniques are used for optical flow computation. These can be grouped as the following categories[11] : differential methods; frequency based methods; correlation based methods; multiple motion methods and temporal refinement methods. The boundaries between each class of methods are not always clear. For instance, both phase based and feature based matching are incorporated in a unique approach in [12], while in [13], a differential scheme is used on time varying edge maps. The former is classified as a phase based method while the latter as a differential method. Noise in the data (the captured images or the modelling technique used) [14] causes the different optical flow computational techniques to give biased flow estimates [15]. The problem of robustness of the different optical flow methods has been particularly investigated in [16, 17], in which different techniques have been examined. The evaluation results showed that the phase-based technique presented in [18] and the differential technique of Lucas and Kanade (KL) [19] produced the more accurate results in overall. Robustness of the differential KL has been also highlighted in [20, 21], in which more complex synthetic image sequences with different techniques investigated in [16, 17] have been re-examined (with exclusion of the phase-based methods).

2.2.5 Background segmentation post-processing

To fit with real-time processing requirements, and due to its reduced computational cost, background subtraction based methods are commonly used in embedded systems. Therefore, after the video source is processed by one of the

background separating methods, the foreground objects (Regions of Interest) are subtracted from the background model. The results of this operation are forwarded to the filtering and thresholding unit where video acquisition noise is removed. By applying a thresholding filter to the processed data, the image is binarised and the amount of information is significantly reduced. These groups of pixels refer to the detected objects. However, to allow for tracking or classification, this data needs further processing.

2.2.6 Connected-component labelling

The connected-component labelling (CCL) algorithms analyse binary images in order to distinguish disjoint groups of pixels (objects) and assign individual labels to them. Labelled objects are further processed to calculate their features which are used by tracking algorithms, such as position, width, height or centre of gravity (centroid)[22]. The CCL is an operation where groups of connected pixels (connected components) are classified as disjoint objects with unique identifiers (labels). There exist different algorithms that deal with the problem of connected component labelling, this includes the multiple scan algorithm [22], the parallel processing algorithm [23], the contour tracing algorithm [24], the single pass algorithm [25] and the two pass algorithm [26].

The multiple scan algorithm [22] is an iterative algorithm that does not require any additional storage for label equivalences. It works by multiple forward, and backward raster scan passes through the image until no label change occurs. Although this algorithm was designed for systems with limited memory resources (low resolution images), its performance is related to the size and the complexity of the binary image which is hard to predict. Therefore, such algorithms are not suitable for real-time video processing. The Parallel processing algorithm proposed in [23], requires higher computational cost as these are principally designed

for parallel processing platforms. Thus, are neither suitable for ordinary computer architectures nor for embeded systems of reduced computational capabilities.

The contour tracing algorithm [24] which is designed to detect contours of the object and also to fill-in interior areas, works through random access to all the image pixels. Therefore, longer execution times are required. The single pass algorithm [25], which is relatively new, was developed specifically for labelling connected components in streaming data systems. The labelling step is performed in a single scan while data is streamed to the system.

In our work we used the two pass algorithm [26] because of its low computational cost, although it requires relatively higher memory usage. It is often referred to in the literature as the classical algorithm. Its key feature is the constant number of passes (two) through the binary image. The general concept is to assign preliminary labels while new foreground pixels are appointed during the initial scan. Once label ambiguity is encountered, the lower label is assigned and the equivalence table (ET) is updated. At the end of the scan. the ET is stored. During the second scan all the preliminary labels are overwritten with their equivalences.

2.2.7 Fundamental concepts for image projection

A portion of this thesis is related to video surveillance systems in which it is very important to accurately estimate the position of detected targets from visual data, it is important to define the transformation linking an image to real world metrics and vice versa. The simplest model which describes this transformation is referenced as pinhole perspective projection model[1, 27]. This model assumes that light rays reflected from the scene pass through a small hole punched on a screen.

The transformation between a position (X, Y, Z) into a pixel location (u, v) where the focal length f characterise the distance between the screen and the image is given by this model (Figure 2.3). In absence of blur and distortions the relationship between the 3D and the 2D coordinates gives the perspective transform defined as follows:

$$\begin{cases} u = f \frac{X}{Z} \\ \text{and} \\ v = f \frac{Y}{Z} \end{cases} \quad (2.2)$$

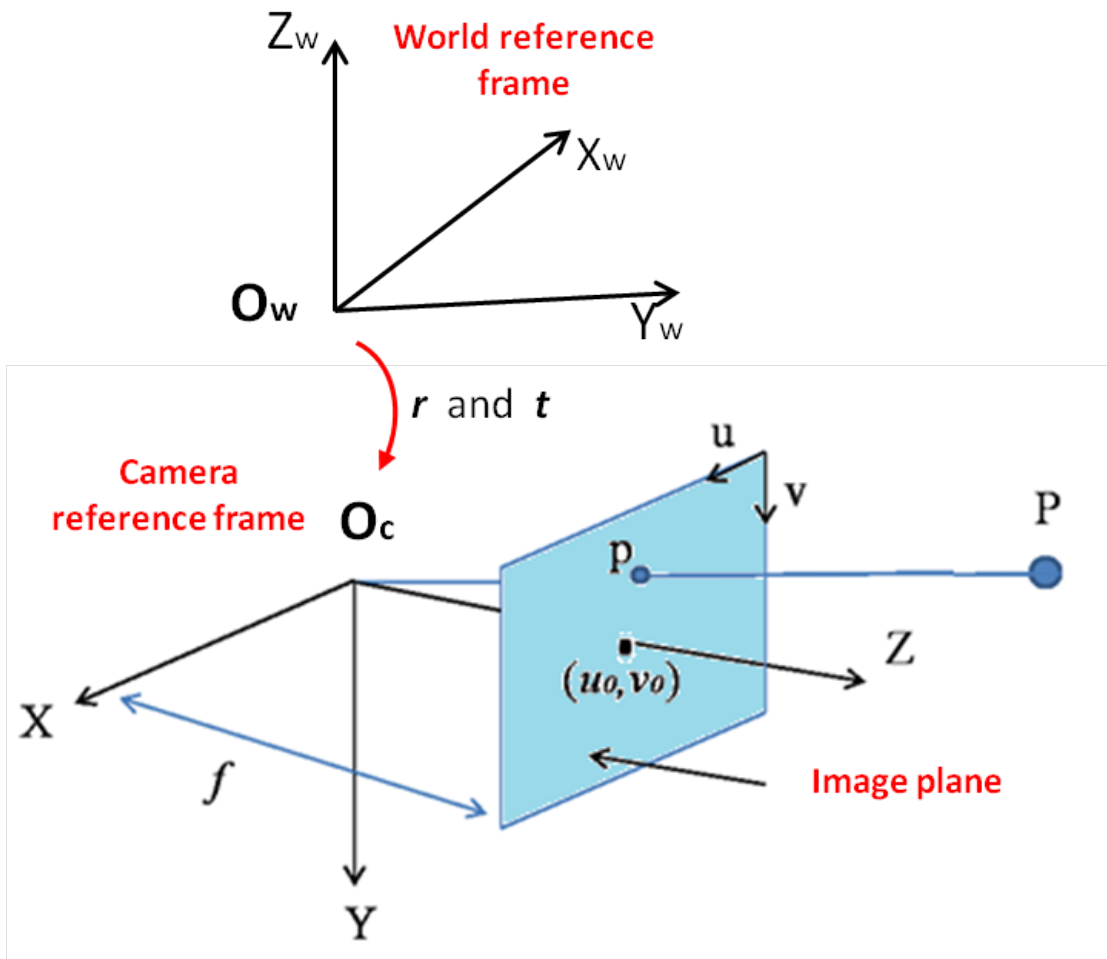


FIGURE 2.3: Camera world reference transformation

where f is the focal length and characterises the distance between the image plane and the projection centre (O_c).

In practice, this model is more complicated as it implies lens distortion, which affects f resulting in an image with distorted corners. In order to have undistorted images, the camera needs to be calibrated, which means estimating the extrinsic and intrinsic parameters.

The intrinsic transformation handles lens distortion and achieves the projection following the pinhole camera model. Optical distortion can be modelled following the radial lens model characterised with two coefficients k_1 and k_2 . Thus, distorted pixel coordinates can be undistorted as follows:

$$\begin{cases} u = u_d(1 + k_1(u_d^2 + v_d^2) + k_2(u_d^2 + v_d^2)^2) \\ v = v_d(1 + k_1(u_d^2 + v_d^2) + k_2(u_d^2 + v_d^2)^2) \end{cases} \quad (2.3)$$

where $[u_d, v_d]^T$ are the distorted pixel coordinates in the image.

The projection transform is characterised by a matrix K , which links homogenous coordinates of \tilde{p} to its related 3D position P as follows:

$$\tilde{p} = KP \quad (2.4)$$

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & \gamma & u_0 \\ 0 & f_v & u_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (2.5)$$

The matrix K is composed of 5 parameters: f_u and f_v the focal length respectively in the horizontal and vertical directions, the skewing factor γ and the coordinates of the central pixel $[u_o, v_o]^T$. The central pixel is the position in the image where the optical axis crosses orthogonally the image plane. Generally, the central pixel does not coincide with the image centre.

The relationship between f_u and f_v , is the following:

$$f_u = s f_v \quad (2.6)$$

where s is a scale factor, which is equal to 1 if the image pixels are square. Finally, the skew factor, which represents the angle between the directions, usually fixed to zero because in reality its value is negligible as these two directions are perpendicular.

A rotation matrix r and a translation vector t define the extrinsic transformation. The latter defines the relationship between the camera reference frame and the world reference frame as illustrated in Figure 2.3. Thus, the operation, which moves a point P_w from the world reference frame to its related position P in the camera reference frame is described as:

$$P = r P_w + t \quad (2.7)$$

2.3 Acoustic source localisation

This Section will be devoted to presenting fundamental definitions of acoustics as these are the basis for various applications in acoustic. That are basically detection, localisation recognition, and classification. Most definitions have been internationally standardised and are listed in standards publications [28].

2.3.1 Physics of sound

Sound or noise is the result of pressure oscillations (or variations) in an elastic medium (air, water, solids) and generated by a vibrating surface, or turbulent

fluid flow. Sound propagates in the form of longitudinal (as opposed to transverse) waves, involving a succession of compressions and rarefactions in the elastic medium, as illustrated by Figure 2.4(a).

When a sound-wave propagates in air (the medium considered in this work), the oscillations in pressure are above and below the ambient atmospheric pressure.

Noise can be described as "undesired or disagreeable sound". From an acoustics point of view, noise and sound present an identical phenomenon of atmospheric pressure variations around the mean atmospheric pressure. The difference is greatly related to the context, as what is sound to one person can be considered as noise to another person.

2.3.2 Amplitude, frequency, wavelength and velocity

A pure tone means that sound consists of a single frequency [29]. It is characterised by the following aspects:

The amplitude of a pressure changes: This feature can be described either by the maximum pressure amplitude (P_m), or the root-mean-square (RMS) of the amplitude (p). It is expressed in Pascal (Pa).

The wavelength (λ): It represents the distance travelled by the pressure wave during one cycle.

The frequency (f): It is expressed in Hertz (Hz) and represents the number of pressure cycles in the medium per unit time, or simply, the number of cycles per second. Noise is usually composed of many frequencies combined together. The relation between wavelength and frequency can be seen in Figure 2.4(b).

Obviously, **the period (T)**, is the time taken for one cycle of a wave to pass a fixed point. It is related to frequency by: $T = 1/f$.

The speed of sound wave propagation (c): is the velocity at which sound travels through a particular medium. It can be written as function of f , the

frequency, and, λ , the wavelength as the following: $\lambda = c/f$. The propagation speed of sound in air is $c = m/s$, at $20^\circ C$ and 1 atmosphere pressure. At different temperatures (t) which present low variations from $20^\circ C$ is calculated as:

$$c = 332 + 0.6T_c. \quad (2.8)$$

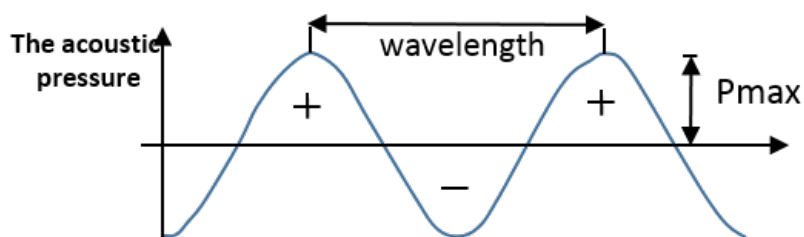
where T_c is the temperature in C° . For estimating the sound speed for any gas, the following expression is used [28]:

$$c = \sqrt{\gamma RT_k/M}(m/s) \quad (2.9)$$

where T_k represents the temperature in K, R is the universal gas constant that has the value of $8.314J$ per *mole* K and M is the molecular weight equal to $0.029 kg/mole$ for air. γ is the ratio of specific heats equal to is 1.402.



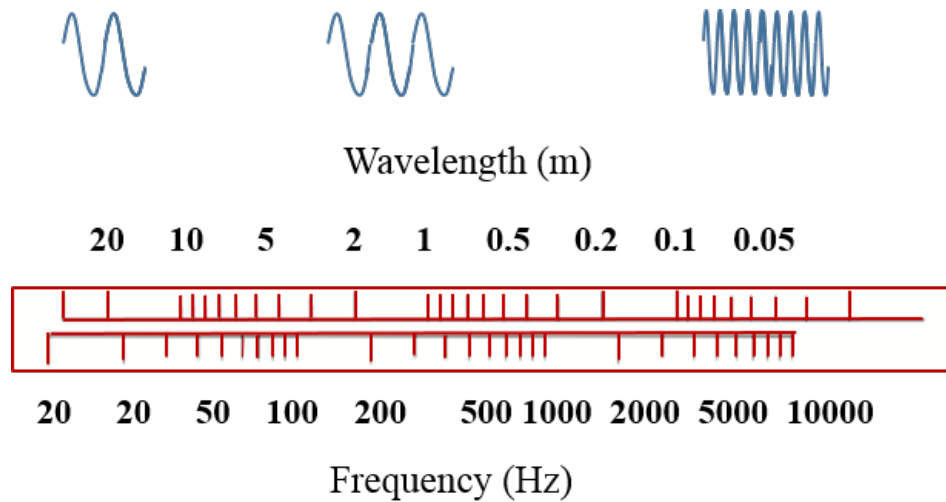
(a) compressions and rarefactions caused in air by the sound wave.



(b) graphic representation of pressure variations above and below atmospheric pressure

FIGURE 2.4: Representation of a sound wave¹

¹Image taken from [30]

FIGURE 2.5: Wavelength versus frequency under normal conditions(air)²

2.3.3 Sound field

It describes a part of the environmental factors affecting the sound wave propagation. More specifically, a sound field is defined by the allowable variation of sound pressure level produced by a loudspeaker in a small space surrounding a reference point [30]. In what follow, we briefly mention the sound field types cited in [28].

- The Free field: defined as the region in space where sound waves propagates free from any form of obstruction.
- The Near field: defined by the region close to an acoustic source where the sound pressure and acoustic particle velocity are not in phase. In this region, the near field is limited to a distance from the source which is equal to about a wavelength of the sound or approximately three times the largest dimension of the sound source [28].
- The Far field: begins at the end of the near field and extends to infinity. Here, the rate of most machinery noise sources attenuation is about 6 dB each time the distance from the source is doubled.

- The Direct field: is described as the part of the sound field where the propagated sound wave does not suffer from any form of reflection caused by room surfaces or obstacles.
- The Reverberant field: is described by the part of the sound field in which the sound wave may experience at least one form of reflection from a boundary of the room or from the enclosure containing the sound source.

2.3.4 Basic Acoustics Features

Since the signal is more stable within a short-time period, a short-term analysis method generally adopted for acoustic signal processing. Usually a step of frame blocking is performed during the processing. At this stage, there may be some overlaps between neighbouring frames to capture subtle change in the acoustic signals. Each frame is the basic unit for audio signal analysis. Within each frame, three most distinct acoustic features can be observed. These are [31]:

2.3.5 Features in the time domain

- **Average energy:** this feature represents the loudness of the audio signal. It is correlated to the amplitude of the signals. The average energy by mean-square values is given by:

$$E = \frac{1}{N} \sum_{n=0}^{N-1} |x(n)|^2 \quad (2.10)$$

where E is the average energy of an audio signal $x(n)$, while N describes the total number of samples in this signal.

- **Zero-crossing rate:** It indicates the frequency of signal amplitude sign change. Estimation of the average zero-crossing is done as follows:

$$ZCR = \frac{1}{2N} \sum_{n=1}^N |sgn(x(n)) - sgn(x(n-1))| \quad (2.11)$$

where $sgn(x(n))$ is the sign of $(x(n))$

- **Silence ratio:** It gives an indication about the proportion of the silence within an audio signal. Silence is described as the interval where the absolute amplitude values are below a given threshold. However, the silence ratio is defined by the ratio between the sum of silent intervals and the total length of the audio signal. Generally, two thresholds are defined: one is used to check if an audio sample is silence and the other is used to determine if those silence samples belongs to a silence interval.

2.3.6 Features in the frequency domain

- **Spectrum:** sound spectrum represents the variation of the signal amplitude with respect to the corresponding frequencies. The spectrum shows the energy distribution across the frequency range.
- **Bandwidth:** it shows the bounds within which sound frequency varies. A simple definition of this feature is the frequency difference between the lowest and highest frequency of the non-zero spectrum components.
- **Harmonic:** in harmonic sound, the spectral components are the multiples of the lowest frequency. This frequency is called fundamental frequency. a well known method used to check if a sound is harmonic compares the frequency corresponding to the dominant components and see if it is a multiple of the fundamental frequency [31].

- **Pitch:** this is a subjective feature which is not only connected to the fundamental frequency. Though in practice, the fundamental frequency is used as approximation of this feature. The most obvious sample point within a fundamental period is often referred to as the pitch mark. The latter is selected as the local maxima or minima of the audio waveform. Reliable identification of pitch is an essential task for some audio applications such the case of text to-speech synthesis.
- **Timbre:** it is an acoustic feature that refers to the 'content' of an audio signals frame. One of the most representative instances is the Mel-frequency cepstral coefficients (MFCCs)[32]. In this representation, the frequency bands are ranked logarithmically in analogy to the response of the human auditory system. The MFCCs performs better than the linear-spaced frequency bands, such as the discrete Fourier transform (DFT) and the discrete cosine transform (DCT) [31] for speech recognition applications.

2.3.7 Methods to analyse audio signals

Different methods are used for acoustic signal analysis, these vary in capabilities and conditions of application.

2.3.7.1 Fourier Transform

The Fourier transform represents a given function in terms of a weighted sum of sine and cosine functions. It is given by:

$$FT(x(t)) = X(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}, \Omega = 2\pi F \quad (2.12)$$

This function is characterised by being impractical to use if we are interested in a specific part of the signal delimited by the interval $t_0 \leq t \leq t_1$.

This method defines the global representation of the frequency content of the signal $x(t)$ over a total period of time in which the signal exists. Also it does not give access to the signal's spectral variations during this interval of time [28]

2.3.7.2 Time-frequency analysis

To address the issue of the Fourier transform, the Short-time Fourier Transform (STFT) has been introduced. It operates by applying the Fourier transform to successive portion of the signal by means of a sliding window of finite size. it is given by:

$$STFT(g(\Omega, b)x(t)) = X_g(\Omega, b) = \int_{-\infty}^{\infty} x(t)g(t - b)e^{-j\Omega t} dt \quad (2.13)$$

where $g()$ is the sliding window. This method was shown in [33] to be efficient for adaptive signal processing. However, once a particular analysing window has been chosen, it cannot be changed again until the end of the entire analysis procedure.

2.3.7.3 The continuous wavelet transform

In order to yield high resolution, specially to analyse low-frequency signals, which is a compatible with the situation in acoustic sensor motes in Wireless Sensors Networks [34, 35]. The Wavelet transform was shown to provide a better trade-off between time and frequency resolutions than the fixed length windows used in the STFT. It is defined as:

$$X_{\psi}(a, b) = (x(t), \psi_a b(t)) = \int_{-\infty}^{\infty} x(t)\psi_{ab}(t)dt \quad (2.14)$$

where ψ_{ab} is a continuous affine transformation of the mother wavelet $\psi(t)$:

$$\psi_{a,b} = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \quad (2.15)$$

$a \in R^+$ and $b \in R$ represent the scaling and the translation parameters respectively. The signal $x(t)$ can be retrieved from its CWT if the constraint:

$$\psi(0) = \int_{-\infty}^{\infty} \psi(t)dt = 0 \quad (2.16)$$

One drawback that affects the continuous wavelet transform is its computational cost due to high level of redundancy.

2.3.8 The generalised cross correlation

There exists different techniques for signal processing that can be used to estimate the time delays between two copies of a signal recorded by a pair of microphone. The most popular is the generalised cross correlation (GCC) with Phase Transform (PHAT) weighting [36]. This technique measures the similarity between one signal and a time delayed version indicates how much time delay between the two versions. The GCC between two signals $S_1(t)$ and $S_2(t)$ is defined as:

$$R_n(t) = \frac{1}{2\pi} \int_0^{2\pi} \frac{S_1(\omega)S_2^*(\omega)}{|S_1(\omega)S_2^*(\omega)|} e^{j\omega t} d\omega \quad (2.17)$$

Where $S_1(\omega)$ and $S_2(\omega)$ represent the Fourier transforms of $S_1(t)$ and $S_2(t)$, respectively. R_n is their weighted cross correlation. Subsequently, the most likely time difference of arrival (TDOA) is equivalent to:

$$TDOA = \underset{\tau}{\text{ArgMax}} (R_n(\tau)) \quad (2.18)$$

2.3.9 Acoustic source localisation in wireless sensor networks

The wireless sensor networks (WSN), is the most known form of Distributed sensor networks. It consists of spatially distributed autonomous nodes, to which sensors are connected. These nodes are used for a wide range of applications

where, often, the main goal is to monitor a specific phenomenon [37]. WSN technology is witnessing a rapid development. A significant amount of research in this area focuses on solutions related to its design, communication issues, energy architecture and extended sensing consumption. However, a neglected part of the efforts focuses on customising feasible applications that can be used in real life applications.

The WSN is built of 'nodes' from a few to several hundreds, where each node is connected to one or several sensors. Each sensor node has typically several parts including [37]: a radio transceiver with an internal antenna or connection to an external antenna, a micro controller, an electronic circuit for interfacing with the sensors; in addition to an energy source, usually a battery or an embedded form of energy harvesting.

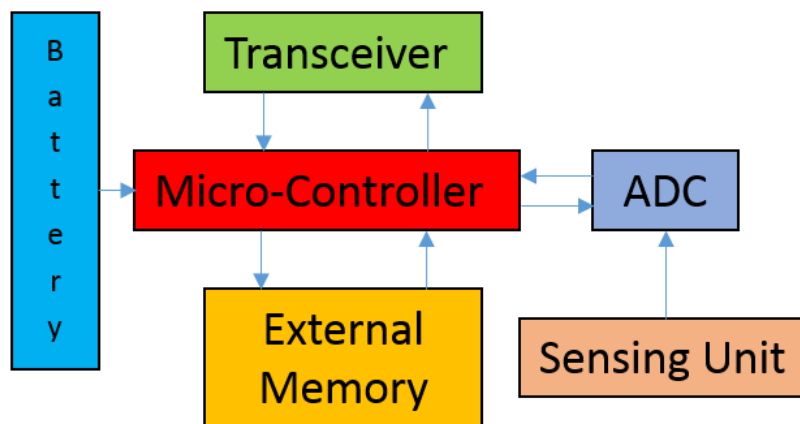


FIGURE 2.6: Sensor node architecture

The acoustic source localisation, as a subject of this study, is one of the most important monitoring tasks which has recently caught the attention of WSN researchers. Its aim is the localisation of an acoustic source from received signals. The usefulness of such application is obvious, especially for security and surveillance purposes

2.3.10 Architectures of distributed systems for acoustic localisation

Although a distributed sensors system is composed of a collection of sensors it appears as a single coherent system to its users.

These distributed systems can further be divided into either centralised or decentralised categories. The centralised systems is a set of remote subsystems are led by a master controller. However, in the decentralised systems, the truly distributed system, has no master controller, and each subsystem manages a portion of the entire system [38].

2.3.10.1 Decentralised systems

This architecture is used in large-scale sensor networks where centralised data processing is not desirable because of the excessive communication and computational complexity it requires [38]. The adopted solution in this situation is to cluster the sensors into groups that collaborate to locate sources. A large number of sensors are required for acoustic source localisation to not only enhance the accuracy but also to increase the robustness of the overall system. Moreover, this class of algorithms has shown its efficiency in terms of both energy and bandwidth usage.

2.3.10.2 Centralised systems

An architecture corresponding to a centralised system is of a common use in this technology. For this architecture, the computations are done by a single controller, which yields a relatively simpler overall design. The sensing nodes are also fairly basic and are often composed of only one sensor or an actuator with minimal to no computing power at all. Thus, these are easier to develop with reduced

cost and development time. However, a major drawback for this architecture lies in the fact that they are not scalable (outputs cannot not be adapted to change in the inputs). Therefore, a controller on a distributed centralised system with hundred nodes, for example, might perform with much less efficiently than a controller supervising only ten. Another drawback to the centralised design is that if the central controller drops off the network for some reason, the system will be useless as this will affect the ability of the nodes to sense or detect their surrounding environment. Their controller will also be unable to collect and archive events that may occur during the downtime.[38]. To detect and localise acoustic sources, this class uses a two-stage approach:

- The first stage is performed at the sensors level where acoustic signals are measured and recorded.
- The second stage, is completed at the central level, in which the collected data are used to calculate the position of the acoustic source.

A great part of the valid data is always available during data collection from a different sensor nodes in this architecture, which will result in a better system performance. However, the overall accuracy is expected to degrade regardless of the optimisation method used for the localisation as part of the data are either missed or corrupted due to the limited network resources. Moreover, this architecture will also increase the communication demands on the system which reduces the lifetime of wireless sensor networks with finite energy resources [38]

2.3.11 Middle-Ware services

The acoustic sources localisation in WSN uses several middle-ware services. If these services are classed in term of importance, the time synchronisation mechanism comes first since precise time synchronisation is crucial in the localisation

application, as sensor nodes need to coordinate their operations and collaborate to complete a complex sensing task [39]. The communication protocol and message delivery are also of great importance for the localisation as the sensors have to send their measurements at approximately the same time and to the same destination. Thus, small latency in sending the obtained measurements renders it useless [40, 41]. Further, sensor node localisation is the base of acoustic source localisation, as the location of the sensor node should be accurate enough for the processing to be meaningful [42].

2.3.12 The Signal Model

Techniques of acoustic source localisation vary according to the sensing and processing capabilities of the systems to be used. In a Distributed Sensors Networks, two main approaches are commonly used. These are either energy based source localisation [43, 44] or time delay of arrival (TDOA) based approach [45].

2.3.12.1 The energy based approach

This technique is motivated by the observation that the sound level decreases as the distance between the sound source and the listener becomes larger. The relation between the sound level and the distance from the sound source enables the estimation of the source position using multiple energy reading at different known source location [44] according to the following formula:

$$y_i = g_i \frac{s}{\|r - r_i\|^\alpha} + e_i, i = 1, \dots, n \quad (2.19)$$

With y_i represents the signal energy measured on the i^{th} sensor ; g_i is the gain factor of the i^{th} acoustic sensor; ; $\|r - r_i\|$ represents the euclidean distance between the i^{th} sensor given by its Cartesian coordinate r_i and the sound source given by its Cartesian coordinates r ; $\alpha = 2$ is an energy attenuation factor;

n is the number of deployed sensors, e_i represent the background noise, it can be approximated using a normal distribution with a positive mean value μ and variance σ .

2.3.12.2 Time delay of arrival (TDOA)

To this approach, we can also assign the DOA based method. The TDOA based method comes from the observation that the sound wave propagates at constant speed (sound speed) from the acoustic source to listeners. A number of microphones at known positions receive the propagated acoustic source at different times. The modelling of TDOA enables the estimation of sound source position using different techniques. An extensive investigation of this method is given in Chapter 5. The DOA based method rely on the array geometrie. From this known geometry, the signal Direction Of Arrival (DOA) can be obtained by measurement of the time-delays which are estimated for each pair of microphones in the array. A best estimate of the DOA is obtained from time-delays and the array geometry [46].

This page is intendedly left blank

Chapter 3

Visual Object Detection with Spatially Global Gaussian Mixture

3.1 Introduction

One of the main objectives of an Automated Visual Surveillance System (AVSS) is intrusion detection. It aims to automatically determine the presence of new objects in the scene. Development of AVSS applications, in recent years, has been a primary area of research interest. However, the proposed solutions are still constrained by the variability of object appearance, poor quality of images, fast lighting variations and object occlusions. Common visual detection approaches for surveillance systems are based on background subtraction. The latter consists of detecting changes in the scene across the frames by comparing the current image with a reference image of the background. Pixels with significant variations are classified as foreground. The popularity of background subtraction

approaches is due to their reduced computational cost in comparison to feature-based approaches.

However, a known problems with background/foreground subtraction is in maintaining the background model when parts of the background are occluded for long periods of time; illumination changes, which can severely affect the accuracy of detection; foreground objects casting shadows where the shadow might be interpreted as foreground; objects believed to be background could move whilst moving foreground objects could stop for a long time.

In this chapter, we investigate the problem of robust background modelling which is able to handle both static and dynamic backgrounds with reduced computational cost. The contributions of this chapter are as follows:

- (i) We present the GMM based foreground/background subtraction approach for object detection. In addition to its performance, it presents an attractive complexity and computation reduction in comparison to classical GMM based visual detection techniques. The background is modelled by a Spatial Global Gaussian Mixture Model (SGGMM) using image colours (RGB) firstly initiated by [47], as an alternative to pixel-based [48] and region-based [49, 50] GMM models. A support map, which stores SGGMM components to pixels assignments, is generated and used in the segmentation process. The background model is updated by processing new image frames online to capture scene changes;
- (ii) To deal with background motion which makes most foreground/background subtraction techniques deficient. A patch-based pixel uncertainty model is used to augment the background colour SGGMM model in a coarse-to-fine detection strategy;

- (iii) Finally, the presented detection algorithm is evaluated quantitatively and qualitatively using a benchmarked dataset in addition to an implementation and a test in a smart camera sensor network node.

3.2 Related Works

To obtain a robust model of the background, several statistical methods have been proposed in the literature. The first category of these methods models the background pixel by pixel using primitive approaches. Some of them use a single Gaussian distribution such as the running average [51, 52] multiple Gaussian distributions or a nonparametric clustering technique [53, 54] to model the pixel colour/intensity. These methods work efficiently in less dynamic scenes but they suffer from vacillating backgrounds (e.g. swaying trees), moving background elements and illumination changes. To alleviate these issues, more advanced techniques have been introduced. These comprise: the Gaussian Mixture Models, adaptive background learning models, models with advanced statistical approaches, in addition to models that are based on outlier detection using low-rank minimisation models. These techniques are detailed in the following:

3.2.1 Gaussian mixture models

Approaches in this category are based on multi-modal Gaussian distributions as the Gaussian Mixture Models. These models are updated over the acquired images, using the Expectation-Maximisation (EM) algorithm [55] or more accurate Bayesian update procedures [56]. GMMs can cope with sudden light changes and work better than other classical algorithms [57]. Stauffer and Grimson presented an example of GMM based methods in [48]. It uses a similar approach as Friedman and Russel's [58] but it was extended to allow several backgrounds to be modelled simultaneously. This is done to account for critical situations such as

waving trees in windy environments. Stauffer's model[48] has gained popularity because of its ability to deal with slow illumination changes and slow moving objects such as those appearing and disappearing from the scene. However, it struggled against fast lighting changes, camera-shaking effects and cases where parts of the background are occluded for a relatively long time. Hayman and Eklundh [59] extended Stauffer-Grimson's algorithm by using colour variance to deal with the problem of partial occlusion of the background (with foreground objects). Although this method showed some performance improvements, the results were not entirely satisfactory. Alternative improvements to GMM-based approaches aimed to increase their computational speed and their adaptation to background variability. This was achieved either by re-investigating the update equations [60, 61], or by adapting both the number and the parameters of the mixture components for each pixel [62, 63]. While vacillating backgrounds and background elements that are moving are better interpreted, fast illumination changes that received considerable attention remain challenging. Adaptive schemes based on colour-invariant principle under varying illuminations for motion detection were proposed in the literature [64–67]. In [64], the background colour was modelled by a single adaptive Gaussian distribution, while a Gaussian mixture is used to model multi-coloured foreground objects (one Gaussian for each colour of the object). Multiple background colours modelling by a mixture of adaptive Gaussians for each pixel was presented in [65]. In that work, an image was represented by the colour chromaticity, which is robust to fast illumination changes in outdoor environments. The mentioned models use only colour or intensity information for background segmentation. Approaches that are based on correlation measurement of pixels over a fixed-size neighbourhood (i.e. patches) are promising. Integrating the pixel spatial location with the colour to model homogenous regions of the background using a Spatial-Colour Gaussians Mixture Model (SCGMM) [49, 50] falls into this category. Each image pixel is classified as foreground or background depending on the classification of the region distribution. The SCGMM model can lead to a better segmentation if the foreground

objects locations are known a priori. However, as the objects may appear anywhere in the scene, using spatial-colour distribution to model the background and/or foreground is often not enough. Moreover, using high-dimensional feature vectors to describe each pixel drastically increases the computational requirements. As the gradient value is less sensitive to lighting changes and is able to derive an accurate local texture difference measure, the authors in [68, 69] have generalised the spatial GMM model to include both colour and gradient features.

3.2.2 Adaptive background learning models

Neural networks and fuzzy logic systems are parameterised computational nonlinear algorithms for numerical processing of data. In these systems, the knowledge is acquired through a learning process and is stored in the internal parameters (weights). Based on these systems, different background models that deal robustly with fast illumination changes have been proposed. Image patches with artificial neural networks have been investigated in [70, 71] and [72]. Maddalena and Petrosino [70] introduced a 2-D flat grid of neurons that yields to a representation of training samples for a reduced image dimensionality. This representation enables the preservation of the topological neighbourhood relations of the input patterns. The nodes are represented as a combination of a weighted linear values of pixels at the input. Therefore, each node is represented by a weight vector. The identification of moving pixels is performed based on the closeness to the weight vector using a distance measure. This method presents some limitations regarding its fixed network structure in terms of the number of neurons which have to be defined in advance. In addition, they lack of hierarchical relationships representation among the inputs. To resolve these issues, Palomo et al. [72] proposed the use of a growing hierarchical neural network. This hierarchical network is divided into layers. Each layer is composed of a number of single self-organising

neural networks with adaptive structures that are determined during the learning process according to input data. Experimental results using this structure showed a good performance in the case of illumination changes. In the same context, Huang and Do proposed a solution based on multi-background generation [71]. It generates a flexible probabilistic model through an unsupervised learning process to determine the property of a background. Moving object detection is performed by estimating the output of an energy function for each block of pixels. Other recent developments such the one proposed in [73], adopted a fuzzy C-means clustering model which uses fuzzy colour histograms as feature. This model enabled the attenuation of the colour variations generated by background motions while still highlighting moving objects. Though this approach delivers satisfactory results with dynamic backgrounds, its main drawback lies in its high sensitivity to the range of proximal pixels. Finally, we have to note that the main difficulty in implementing methods based on neural networks and fuzzy logic is in their higher computational load and the number of clean frames for the training phases.

3.2.3 Advanced statistical models

This category comprises robust models based on advanced statistical analysis, which is applied to pixel patches within the acquired images. In [74], Cheng and Huang proposed a temporal difference based method, which adopts a Laplacian distribution model to check for the presence of moving objects in different image blocks. It also uses an adaptive background model and illumination variation mechanism with a training procedure. Computing the binary object detection mask using a suitable threshold value performs a motion extraction task. Although this approach shows some level of robustness to environmental changes, its application often results in incomplete detection of the moving objects shape. This is especially true when objects, which are motionless or with feature limited

mobility, are present. In [75], Huang proposed a background subtraction method using a model based on the selection of suitable background candidates. The method uses an Alarm Trigger (AT) module, which relies on a block-based entropy evaluation (using morphological operations) to detect the pixels of moving objects within the regions designated as belonging to feature objects. Similarly, Guo et al [76] proposed a multilayer codebook-based technique. It works through the combination of a multilayer block-based strategy and the adaptive feature extraction from blocks of various sizes to remove most of the non-stationary (dynamic) background. Thus, the processing efficiency increased. Though it is efficient in some scenarios, a refining step of the result is still required. The latter uses a pixel-based classification scheme to identify pixels as either foreground, shadows or highlights.

3.2.4 Low-rank minimisation models

Low-Rank minimisation (LRM) based methods show their usefulness in many data mining applications. However, their performance can seriously degrade due to outliers. To resolve this issue, extensive investigations have been conducted to develop robust matrix factorisation methods. This is done by formulating the problem of robust LRM as a constrained matrix approximation problem with constraints on the rank of the matrix, the cardinality of the outlier set, in addition to incorporating outlier structural knowledge. Using this approach, Zhou and al. proposed a method based on the assumption that the underlying background images are linearly correlated [77]. Therefore, by finding a Low-rank matrix that approximates the vectored video frames, moving objects can be detected from the outliers. This presented method works in batch mode only. Thus, a further investigation is required to adapt it for real-time scenarios. In [78], Xiong et al proposed a (Direct) Robust Matrix Factorisation (DRMF) approach. It presumes the existence of some arbitrary outliers in a small portion of the matrix. Then, to

get a reliable estimation of the true Low-rank structure of this matrix and to identify the outliers, the latter should be removed from the model estimation. This can be done as long as these outliers are present in a reduced number. In addition, and for computational acceleration, the authors proposed to use a partial SVD algorithm. Similarly, Wang et al [79] proposed a Probabilistic Robust Matrix Factorisation (PRMF) method. This was formulated with a Gaussian prior and Laplace error, which correspond to an l_1 loss and an l_2 regularised respectively. For the model learning process, a parallelizable Expectation Maximisation (*EM*) algorithm was introduced. Additionally, an online extension of the algorithm for sequential data was provided to offer further scalability. The performance of the PRMF is comparable to classical robust matrix factorisation methods. However, in terms of accuracy it performs better for large data matrices. The methods that fall into this category are promising and can perform well in some particular scenarios. However, intensive investigations are still required to resolve issues regarding their reduced overall accuracy and increased computational costs.

3.3 Description of the spatially global gaussian mixture model

A basic assumption for moving objects detection using a static camera, is that the background slowly changes in relation to the moving objects in the scene. Each image pixel value is represented in feature space by a vector $x = [R, G, B]^T$. The scene background is represented by a spatially global Gaussian mixture model of n Gaussians in 3-dimensional colour space as follows

$$p(x) = \sum_{i=1}^n w_i g(x, \mu_i, \Sigma_i) \quad (3.1)$$

where μ_i and Σ_i are respectively the spatial mean vector and covariance matrix of the i^{th} distribution. w_i is an estimate of the weight which reflects the likelihood that the corresponding distribution accounts for the image colour and satisfies the criteria $\sum_{i=1, \dots, n} w_i = 1$. Each Gaussian distribution $g(x, \mu_i, \Sigma_i)$ of the mixture is defined as:

$$g(x, \mu_i, \Sigma_i) = \frac{1}{2\pi^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)(\Sigma_i^{-1})(x-\mu_i)^T} \quad (3.2)$$

Where $d = 3$. The *RGB* colours are assumed to be independent random variables and the covariance Σ_i is diagonal matrix for computation simplicity.

3.3.1 Background model estimation

The mixture components are built at the initialisation phase based on a mean image (I_{mean}), which is computed over a number of frames N taken without any moving objects. We associate a Gaussian probability density function to each pixel of I_{mean} as follows:

$$m_i = \frac{1}{N} \sum_{t=1}^N (x_i(t)) \quad (3.3)$$

$$\sigma_i = \frac{1}{(N-1)} \sum_{t=1}^N [(x_i(t)] - m_i][[(x_i(t) - m_i)]^T \quad (3.4)$$

where $x_i(t)$ is the pixel value in frame t , m_i and σ_i are respectively the mean and covariance of pixel i . For simplicity reasons, the pixel position in the image is represented by one dimension. As each pixel in the image is modelled by a Gaussian distribution, we associate to the mean image I_{mean} a mixture density f of dimension $w \times h$ Gaussian components defined as follows:

$$p(x) = \sum_{(i=1, \dots, w \times h)} \alpha_i f(x, m_i, \sigma_i) \quad (3.5)$$

where w and h are the width and height of the image. $f(x, m_i, \sigma_i)$ is a Gaussian component with a mean and covariance given in equations (3.3) and (3.4), and α_i

is the mixing weight equal to $\alpha = \frac{1}{w \times h}$. The objective is to fit all the components of the mixture f into a reduced mixture g of n components and representative of the background colour. Therefore, the SGGMM estimation can be formulated as a clustering problem so that the set of pixels of a region with similar colour are fitted to the same cluster and represented by the same Gaussian component.

To this end, we adopt an adaptive hierarchical clustering algorithm introduced in [80], where clusters can be created and updated adaptively. If new pixels are fitted to existing clusters, the corresponding Gaussian component parameters are updated accordingly. To determine the elements of the reduced mixture g closest to elements of the original mixture f , a distance minimisation criterion between f and g is used and defined as follows:

$$d(f, g) = \sum_{i=1, \dots, w \times h} \min_{k=1, \dots, n} KL(f_i || g_k) \quad (3.6)$$

where $KL(f_i || g_k)$ is the Kullback-Leibler distance between components $f_i = N(m_i, \sigma_i)$ and $g_k = N(\mu_k, \Sigma_k)$ given by:

$$KL(f_i || g_k) = \frac{1}{2} [\log \frac{|\Sigma_k|}{\sigma_i} + Tr(\Sigma_k^{-1} \sigma_i) + (m_i - \mu_k) \Sigma_k^{-1} (m_i - \mu_k) - c] \quad (3.7)$$

with $c = 3$ is the dimensionality of the image features space. Since there is no closed-form solution to this minimisation problem, an iterative approach to obtain a locally optimal solution is proposed which operates through the following steps:

- Step 0: Initialise the mixture with a single component of an arbitrary chosen mean vector μ_0 and a diagonal covariance matrix Σ_0 (i.e. $\mu_0 = m_1$ and $\Sigma_0 = \sigma_1$)
- Step 1: For each pixel of the mean image I_{mean} , compute the KL-distance of the corresponding component f_t regardless of all components $g_k, k = 1, \dots, n$

of the mixture g using equation (3.7).

$$L_k = \{KL(f_i||g_k), k = 1, \dots, n\} \quad (3.8)$$

- Step 2: Determine the minimum of the KL-distance vector, $L_{min} = \text{argmin}_k L_k$.

For this step, two cases arise:

- If $L_{min} \leq r$, where r is a predefined threshold, the component is fitted to the cluster represented by the component corresponding to the minimum KL-distance;
- If $L_{min} > r$, a new cluster is initialised with the component f_t , while n is set to $n = n + 1$;

- Step 3: Update all the clusters by determining the new parameters associated with the representative components g_k of the mixture, as follows:

$$\mu_k = \frac{\sum_{i \in \pi(k)} \alpha_i m_i}{\sum_{i \in \pi(k)} \alpha_i} \quad (3.9)$$

$$\Sigma_k = \frac{1}{\sum_{i \in \pi(k)} \alpha_i} \sum_{i \in \pi(k)} \alpha_i \left(\sigma_i + (m_i - \mu_k)(m_i - \mu_k)^T \right) \quad (3.10)$$

$$\beta_k = \sum_{i \in \pi(k)} \alpha_i \quad (3.11)$$

Where β_k is the weight, μ_k is the mean, Σ_k is the covariance matrix and π_k is the set of pixels assigned to the cluster represented by the SGGMM component g_k .

- Step 4: Compute the new minimisation distance $d(f, g)$ using equation (3.6) and compare it to that of the previous iteration. Repeat steps 1, 2 and 3 until the minimisation distance difference is kept less than a user-defined $T_{C_{map}}$ threshold.

Given the set of the mixture components g_k , an observed pixel value is assigned to the component with the maximum posterior probability. Thus, a support map C_{map} is built to store the current component assignment for each pixel,

$$c = \text{Arg} \max_{j=1, \dots, n} \log(p(x|g_j)) \quad (3.12)$$

where x is the pixel value. This support map, which is obtained offline from the SGGMM model, will be used during the online image frame processing to perform segmentation. It is updated according to the SGGMM update step.

3.3.2 Background model update

The SGGMM model is updated using pixels segmented to the background. By having a background model from the previous frame I_{t-1} (at instant $t - 1$), and since the background is continuously changing, we seek to update it over the next image frame I_t . The proposed update algorithm operates through the following steps:

- Step 1: Firstly, with the use of the support map, the joint likelihood of the whole image is determined as the following:

$$p(I_t | g) = \prod_{I_t} \sum_{i=1}^N \omega_i g(x_t, \mu_i, \Sigma_i) \quad (3.13)$$

The defined joint likelihood of the image will be used to adaptively define the pixel maximum likelihood.

- Step 2: Pixel likelihood is determined regardless the SGGMM component obtained from the support map, this is done according to the following:

$$L_k = \omega_k g(x_i, \mu_k, \Sigma_k) \quad (3.14)$$

This likelihood value is constrained to exceed an adaptive user defined threshold T_{Cmap} , chosen depending on the joint image likelihood, that is, $T_{Cmap} = \gamma p(I_t | g)$ where $\gamma \in [0, 1]$ is a defined constant. In this case, two cases arise:

- if $L_k < T_{Cmap}$, the pixel is maintained in the same cluster represented by the component g_k ;
- if $L_k > T_{Cmap}$, we determine the pixel likelihood vector regardless of all the SGGMM components as following:

$$L_k = \{p(x_i | g_k), k = 1, n\} \quad (3.15)$$

Then, we determine the maximum value of the likelihood vector, $L_{max} = \text{Arg max}_{k=1, \dots, n} L_k$. If the maximum likelihood exceeds the threshold T_{Cmap} , the pixel is fitted to the Gaussian component with the maximum likelihood. However, if the maximum likelihood falls below the threshold T_{Cmap} , then a new Gaussian component is initiated with its mean equal to the pixel value and covariance matrix set to $\text{diag}([0.0001 \ 0.0001 \ 0.0001])$. After processing the entire image, if an existing component does not match any pixel it will be eliminated.

- Step 3: The Gaussian component parameters (mean, variance and weight) of the new SGGMM background model are updated as follows:

$$\mu_k = \frac{\sum_{i \in \pi_k} p(x_i | g_k) x_i}{\sum_{i \in \pi_k} p(x_i | g_k)} \quad (3.16)$$

$$\Sigma_k = \frac{\sum_{i \in \pi_k} p(x_i | g_k) (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i \in \pi_k} p(x_i | g_k)} \quad (3.17)$$

$$w_k = \frac{\sum_{i \in \pi_k} p(x_i | g_k)}{\sum_{k=1}^N \sum_{i \in \pi_k} p(x_i | g_k)} \quad (3.18)$$

where π_k is the set of pixels assigned to the cluster represented by the SGGMM component g_k , while $p_i(x_i, g_k)$ is the pixel likelihood.

Once the adaptation is performed, a new support map is generated. This background SGGMM update approach resolves the issues of adaptation to low illumination changes and background motion. Additionally, objects being stationary for a little moment (e.g. left baggage, parked car), can easily be incorporated in the background. Furthermore, if an object is stationary enough to be modelled as a background region (e.g. a moved car in a parking, or a baggage left for a long enough time) and then it moves, the background model is updated rapidly to include the moving object. Figure 3.1 illustrates the different components of an SGGMM model (Gaussian mixtures models $((w_i, \mu_i, \Sigma_i)_{i=1, \dots, k})$ with their indices i stored in a support map Cmap. Each cell in C_{map} contains the Gaussian index to which the corresponding pixel in the mean image I_{mean} belongs to. I_{mean} is the mean of the first captured images.

3.4 SGGMM Based Pixel Location Uncertainties Approach

Although the SGGMM can handle, reasonably accurately, problems of moving objects segmentation with static backgrounds, issues are still present in the case of dynamic backgrounds and camera jitter (Figure 3.2). This is because pixels positions are generally affected by some uncertainties. To account for them, the pixel location in the image space is considered as a random variable within a patch. The patch is a region centered at the pixel location (i, j) and composed of (P^2) pixels, where P is chosen according to the maximum displacement of pixels. The proposed SGGMM background model based on RGB colour as presented in Section 3.3, is modified by augmenting the feature vector to include

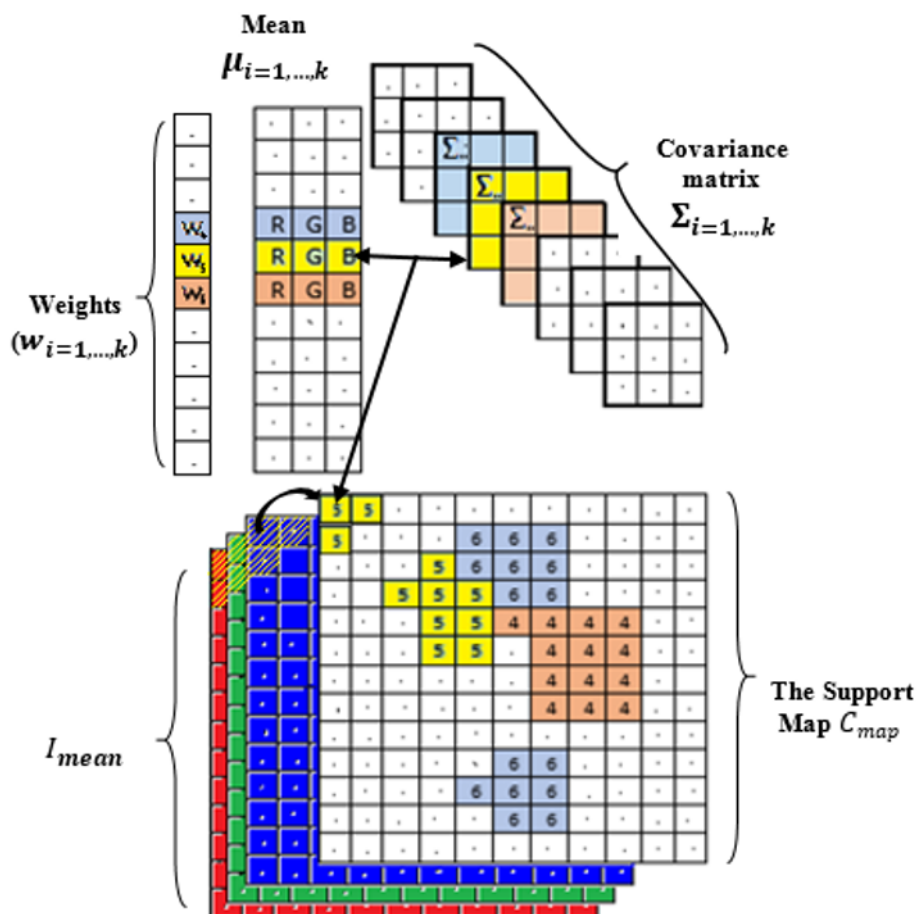


FIGURE 3.1: Different components of the SGGM based model

pixel displacement uncertainties with RGB colour. This leads to what we coin a SGGM based pixel uncertainty approach (Figure 3.3). In this section, we compute the pixel location uncertainties within a neighbouring region using grey level derivatives. A similar approach is used in optical flow methods [81, 82].

3.4.1 Computation of pixel displacement uncertainties

The true location of a pixel in the background image B is assumed to be (i, j) , its corresponding position is (\bar{i}, \bar{j}) in the current image (given within a patch). The pixel location uncertainty components $x = \bar{i} - i$ and $y = \bar{j} - j$ are evaluated in the corresponding patch by minimising the sum function of squared gradients

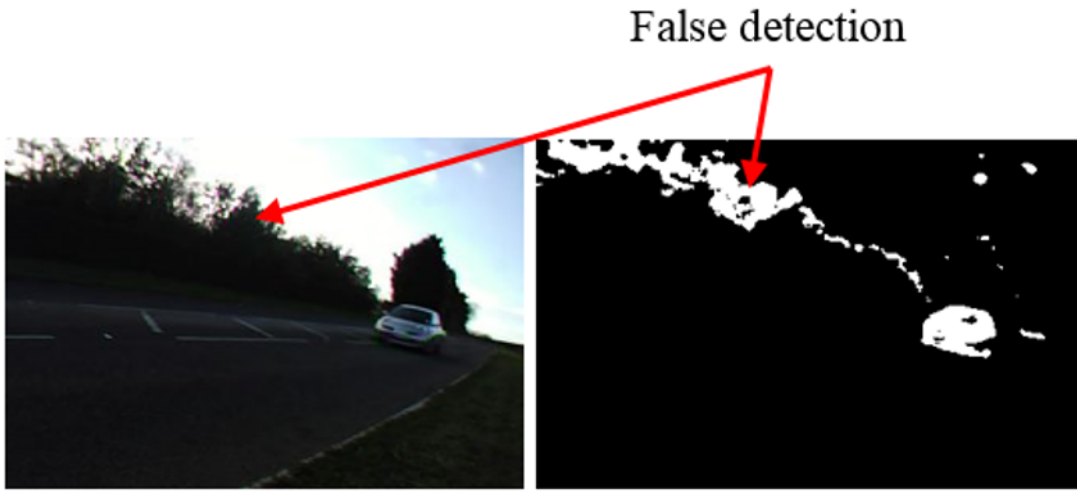


FIGURE 3.2: Effect of the dynamic background on accurate detection

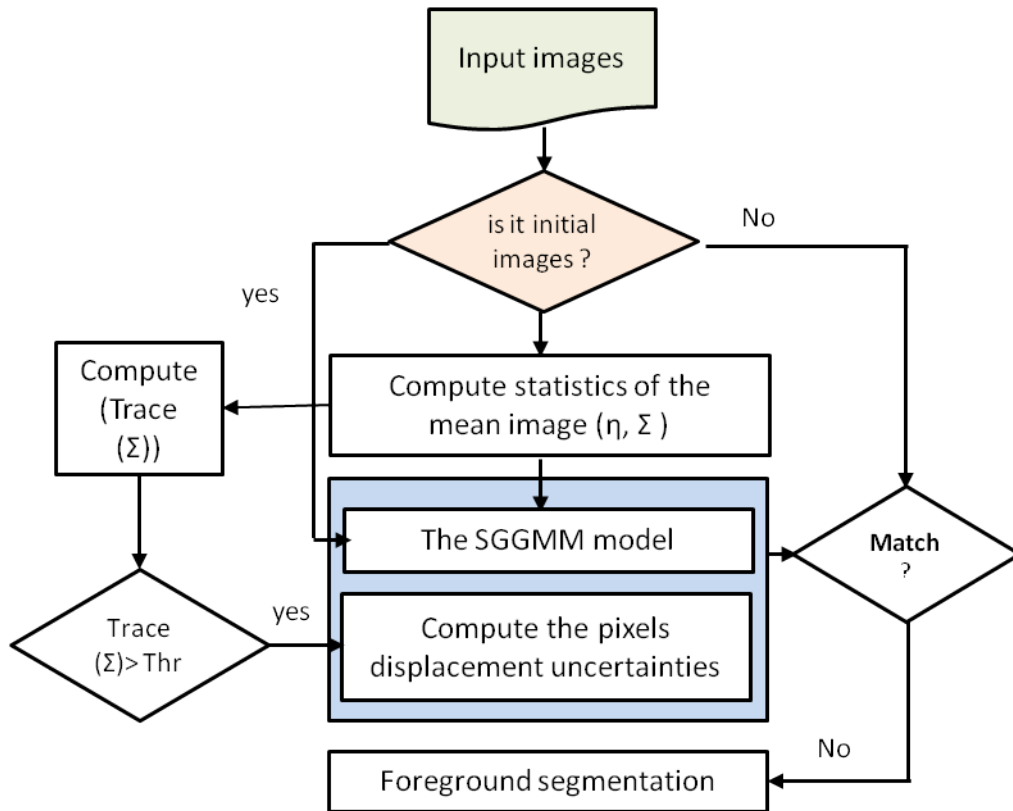


FIGURE 3.3: Background subtraction approach based on pixel uncertainties

between the background image B and the current image I_t as follows:

$$J(x, y) = \sum_{i,j \in P} w_{ij} (I_t(i+x, j+y) - B(i, j))^2 \quad (3.19)$$

where P is the square patch size and w_{ij} is a weighting function describing the pixel proportion in the patch area such that $0 \leq w_{ij} \leq 1$. More importance are given to the central pixels that is why an exponential function w_{ij} of the following form is chosen to assign the higher weight:

$$w_{ij} = e^{-\frac{(i^2+j^2)}{\sigma^2}} \quad (3.20)$$

where $-P/2 \leq (i, j) \leq P/2$ are pixel coordinates in the patch, and σ is a constant, which determines the proportionality rate of pixel gradient over the patch. Theoretically, the measurement process of pixel grey scale generates a measurement error that is used to evaluate the pixel location uncertainties. If we approximate the shifted signal $I_t(i+x, j+y)$ by a first order Taylor series, and assuming that the total derivative of the image intensity function is zero at each position in the image, the error function $J(x, y)$ can be written as follows:

$$J(x, y) = \sum_{i,j \in P} w_{ij} \left(\begin{bmatrix} I_x(i, j) & I_y(i, j) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - B(i, j) \right)^2 \quad (3.21)$$

and

$$\nabla J(x, y) = \sum_{i,j \in P} w_{ij} \left(I_t(i, j) - B(i, j) + \begin{bmatrix} I_x(i, j) & I_y(i, j) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right) \begin{bmatrix} I_x(i, j) \\ I_y(i, j) \end{bmatrix} \quad (3.22)$$

where I_x and I_y are the partial derivatives of the pixel value with respect to the position. By setting the derivative of the error function in Equation (3.22) to

zero, we obtain:

$$\sum_{i,j \in p} w_{ij} (I_t(i, j) - B(i, j)) \begin{bmatrix} I_x(i, j) \\ I_y(i, j) \end{bmatrix} = - \left(\sum_{i,j \in p} w_{ij} \begin{bmatrix} I_x(i, j) & I_y(i, j) \end{bmatrix} \right) \begin{bmatrix} I_x(i, j) \\ I_y(i, j) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (3.23)$$

Thus, the pixel uncertainties can be expressed as follows

$$d = -M^{-1}b \quad (3.24)$$

From Equation 3.23 and 3.24, the displacement vector d , the vector b and the matrix M correspond to:

$$d = \begin{bmatrix} x & y \end{bmatrix}^T \quad (3.25)$$

$$b = \sum_{i,j \in p} w_{ij} (I_t(i, j) - B(i, j)) \begin{bmatrix} I_x(i, j) \\ I_y(i, j) \end{bmatrix} \quad (3.26)$$

$$M = \sum_{i,j \in p} w_{ij} \begin{bmatrix} I_x^2(i, j) & I_x(i, j)I_y(i, j) \\ I_x(i, j)I_y(i, j) & I_y^2(i, j) \end{bmatrix} \quad (3.27)$$

3.4.2 Statistics computation of pixel displacement uncertainty

The spatial derivatives are considered uncertain due to image imperfections caused by lighting changes and low contrast regions. Consequently, the pixel position displacement computation is considered as a probabilistic problem that takes these uncertainties into account. Therefore their statistics should be estimated. The aim at this stage is to capture the nature of this uncertainty distribution. The latter is modelled as normally distributed random variables with a mean m_{xy} and

a covariance Σ_{xy} estimated as follows [83, 84]:

$$m_{xy} = [E(x), E(y)]^T \quad (3.28)$$

$$\Sigma_{xy} = \begin{bmatrix} E(x^2) & E(xy) \\ E(xy) & E(y^2) \end{bmatrix} \quad (3.29)$$

where $E(\cdot)$ is the expectation operation. If the pixel position uncertainties (x, y) are affected with random perturbations $(\Delta x, \Delta y)$, Equation (3.22) can be rewritten as follows:

$$\Sigma_{i,j \in p} w_{ij} \left(I_t(i, j) - B(i, j) + \begin{bmatrix} I_x(i, j) & I_y(i, j) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right) = -\Sigma_{i,j \in p} w_{ij} \begin{bmatrix} I_x(i, j) & I_y(i, j) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad (3.30)$$

We write the right hand side of equation (3.30) as the multiplication of a vector L and a noise vector ϵ with mean m_ϵ and covariance matrix Σ_ϵ .

$$\Sigma_{i,j \in p} w_{ij} \left(I_t(i, j) - B(i, j) + \begin{bmatrix} I_x(i, j) & I_y(i, j) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right) = L\epsilon \quad (3.31)$$

By performing a variable change $\epsilon' = \epsilon - m_\epsilon$, the least squares estimate of the pixel displacement uncertainties is equivalent to the minimisation of the following cost function:

$$e = \left(\begin{bmatrix} x \\ y \end{bmatrix} + M^{-1} \left(b - L^T m_\epsilon \begin{bmatrix} I_x \\ I_y \end{bmatrix} \right) \right)^T L^{-T} \Sigma_\epsilon^{-1} L^{-1} M \left(\begin{bmatrix} x \\ y \end{bmatrix} + M^{-1} \left(b - L^T m_\epsilon \begin{bmatrix} I_x \\ I_y \end{bmatrix} \right) \right) \quad (3.32)$$

Where L^{-1} is the pseudo inverse of matrix L so that $L = L^T(LL^T)^{-1}$. By taking into account the distribution of the cost function e that may be assimilated to x^2

distribution, the pixel position uncertainties (x, y) are considered as a Gaussian distributed random variable with mean vector and covariance matrix derived as follows:

$$\Sigma_{xy}^2 = L^T \Sigma_\epsilon L M^{-1}, m_{xy} = -M^{-1} \left(b - L^T m_\epsilon \begin{bmatrix} I_x \\ I_y \end{bmatrix} \right) \quad (3.33)$$

The distribution parameters of the pixel uncertainties given in Equation (3.33) are fully derived from the spatial derivatives and the background SGGMM model by considering the pixel colour distribution we are looking at from the support map (3.12). There is no need to explicitly update the parameters of the pixel uncertainty distribution since the background SGGMM model, from which they are computed from, is updated at each frame.

3.4.3 Using pixel uncertainty in background subtraction

After modelling the background using the SGGMM model in RGB colour space, with the probability of observing a value $I_{c,i} = [I_r, I_g, I_b]$ at a given pixel i is given by a mixture of Gaussians (Equation 3.1). The next step was to estimate the pixel uncertainties $d = [x, y]^T$ (Equation 3.24), and to compute their statistics (equation 3.33). The resulting feature vector, which include the RGB colour and pixel uncertainties, is considered as a 5-dimensional Gaussian distributed random variable with the following mean vector and covariance matrix.

$$\mu_i = \begin{bmatrix} \mu_{c,i} & m_{xy} \end{bmatrix}, \Sigma_i = \begin{bmatrix} \Sigma_{c,i} & 0 \\ 0 & \Sigma_{xy} \end{bmatrix} \quad (3.34)$$

3.5 Foreground segmentation

The segmentation task is performed in each image frame to classify image regions as either background or foreground, based on the background SGGMM model using spatial and temporal features. Two approaches can be used based on whether the feature vector is containing the RGB colour on its own as presented in section 3.3, or combined with pixel uncertainties. The foreground removal is processed using the obtained support map. Each pixel likelihood, $\log(p_x \| g_{cmap})$, is evaluated in the new frame. If the likelihood is less than a user defined threshold T_{seg} ,

$$\log(p_x \| g_{cmap})_{ij} < T_{seg} \quad (3.35)$$

the pixel is set to the foreground. Otherwise, it is set to the background.

3.6 Performance Tests

3.6.1 Performance evaluation of the SGGMM model

The accuracy of the proposed method is evaluated both quantitatively and qualitatively. Further, we compare results obtained by our technique to the ones obtained using the GMM-Stauffer [48], the Decolor method based on the low-rank minimisation approach [77] and the KDE method [54]. These methods have been tested using optimal parameters designed by their authors for the same datasets. The GMM-Stauffer method is initialised with $k = 3$ number of Gaussians while the learning rate is fixed to $\alpha = 0.001$. The Decolor method is used with the following parameters: the convergence precision is set to $tol = 1e - 4$, the desired rank of the estimated low-rank component is $k_1 = 4$ (equivalent to 20 frames) and the constant for controlling the strength of smoothness is $\lambda = 5$. The KDE method uses a sample size equal to $n = 100$ frames. Firstly, the SGGMM

background models for all considered scenarios (datasets) are estimated. These datasets [85], which are representative of static and dynamic scenes, provide testing image sequences and their corresponding ground truths. Image samples of these data are shown in Figure 3.4. Sequences OFFICE, and PETS2006 are used to evaluate the methods accuracy in static backgrounds. The OFFICE sequence consists of a person moving in an office, whereas, the PETS2006 sequence represents people walking at a train station. The sequences utilised to evaluate the method in dynamic backgrounds cases are CANOE and OVERPASS. The first sequence is of a boat crossing a river while the second one is of a person crossing a bridge.

An analysis of the derived pixels uncertainties statistics has also been conducted to ensure that these are adequately describing the pixel location uncertainties. Indeed, we analysed the mean value and covariance information of the different points (pixels) corresponding to the test sequence (Figure 3.5). This is achieved by computing pixel displacement from subsequent background images in the sequence. Their normalised histograms are then derived. Parts of the images where the points are covered with moving objects are ignored. Figure 3.5 shows the background points histograms shape. It clearly shows the suitability of approximating the pixel displacement uncertainties using a Gaussian distribution as proposed in Section 3.4.2 related to the statistical nature of pixel uncertainties.

3.6.2 Quantitative evaluation

The quantitative evaluation involves the use of different criteria. These are the Recall, the Precision and the similarity metrics, along with the F-measure [85, 86]. The Recall (Re) provides the percentage of detected true positives in a comparison to the total number of items in the ground truth.

$$Re = \frac{tp}{tp + fn} \quad (3.36)$$

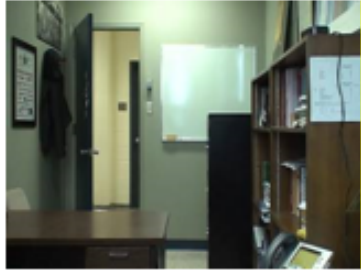


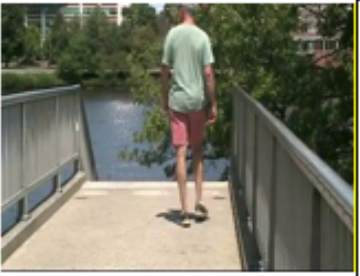
	
Scenes with static backgrounds	
Office	PETS2006
Frame 570 to 2050 360x240	Frame 300 to 1200 576x720
	
Scenes with dynamic backgrounds	
Canoe	Overpass
Frame 800 to 1189 320x240	Frame 991 to 3000 320x240

FIGURE 3.4: Number of test frames for the evaluation of the SGGMM

Where tp is the total number of true positive pixels, fn is the total number of false negative pixels and $(tp + fn)$ represents the total count of pixels representing the detected objects in the ground truth. The Precision (Pr) provides the percentage of detected true positives by comparison with the total count of items in the binary objects mask detected by the algorithm.

$$Pr = \frac{tp}{tp + fp} \quad (3.37)$$

Where fp is the total count of false positive pixels. The total count $(tp + fp)$

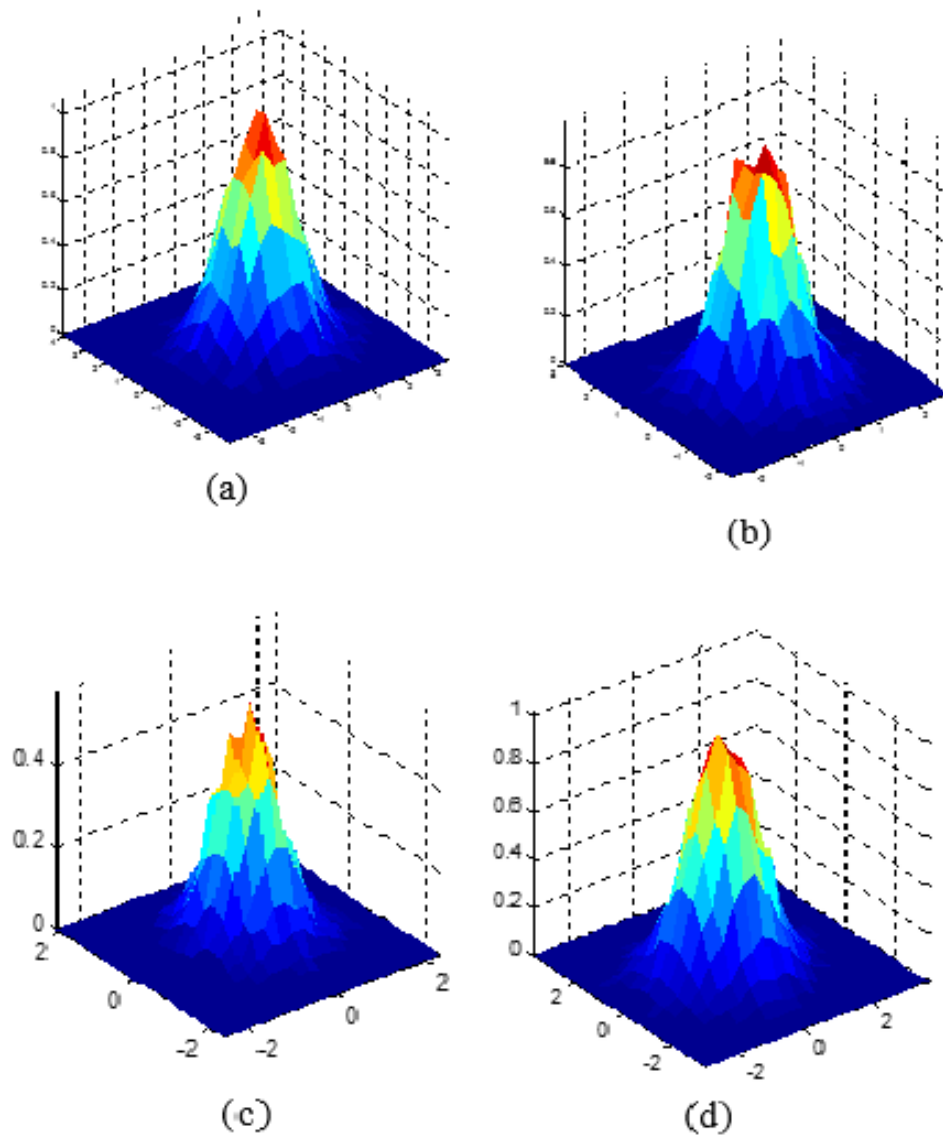


FIGURE 3.5: Pixel location uncertainties normalised histogram of the background points: (a) pixel (100,100) of the first scene, (b) pixel (100,100) of the second scene, (c) pixel (100,100) of the third scene, (d) pixel (100,100) of the fourth scene

represents the detected items in the binary objects mask. The Recall metric evaluates the incorrect association of internal lost pixels to moving objects, while the Precision criterion measures only the incorrect association of false detected pixels. The use of the mentioned metrics alone cannot offer a satisfactory comparison between the different methods. Thus, and in order to deliver a more

Dataset	Evaluation	SGGMM Based Colour Only	SGGMM W.P.U	KDE	GMM-Stauffer	DECOLOR
Office dataset	• Precision	0.8915	0.8978	0.9675	0.7462	0.7192
	• Recall	0.8457	0.9105	0.9054	0.4904	0.7033
	• Similarity	0.7668	0.8916	0.8787	0.2959	0.6324
	• F-Measure	0.8680	0.9041	0.9354	0.84039	0.7115
PETS2006 dataset	• Precision	0.8649	0.8714	0.8283	0.7856	0.7066
	• Recall	0.7304	0.7617	0.7903	0.6596	0.8290
	• Similarity	0.6556	0.7005	0.6791	0.7073	0.6168
	• F-Measure	0.7920	0.8129	0.8089	0.7171	0.7629
Canoe	• Precision	0.8210	0.9505	0.9396	0.8981	0.9874
	• Recall	0.7963	0.8416	0.8314	0.8659	0.6755
	• Similarity	0.7037	0.8063	0.7893	0.7885	0.6697
	• F-Measure	0.8084	0.8927	0.8812	0.8817	0.8022
Overpass	• Precision	0.8102	0.9321	0.6780	0.9190	0.8013
	• Recall	0.7153	0.8583	0.5961	0.8294	0.7433
	• Similarity	0.7281	0.7887	0.4646	0.7729	0.7223
	• F-Measure	0.7597	0.8936	0.6344	0.8719	0.7712

TABLE 3.1: Comparison between the Average Similarity, F-measure, Precision, and Recall values for each method

in-depth evaluation, the accuracy of the proposed algorithm was estimated using two additional metrics. These are the Similarity (S) given by:

$$S = \frac{tp}{tp + fp + fn} \quad (3.38)$$

The second is the F-measure(F), which represents the harmonic means of Recall and Precision. It is given by:

$$F = 2 * \frac{Re * Pr}{Re + Pr} \quad (3.39)$$

The values of the estimated metrics range from 0 to 1, where higher values indicate better accuracy. The quantitative evaluation results for the considered five video sequences and using the above criteria are displayed in Table 3.1. From the latter we can see and conclude that our SGGMM With Pixel Uncertainties (W.P.U)

technique globally outperforms GMM-Stauffer, KDE and Decolor algorithm. The strength of our technique is clear when dealing with dynamic backgrounds. We can also notice that SGGMM based on colour only is providing similar results to SGGMM W.P.U when dealing with static background. On the other hand, when dealing with dynamic backgrounds we can appreciate the contribution of incorporating the uncertainties into the SGGMM.

3.6.3 Qualitative evaluation

Here, we are comparing our technique with the other techniques and also visually showing the effect of including the uncertainty into the SGGMM model.

3.6.3.1 Segmentation using colour based SGGMM

Figure 3.6 and 3.7 show segmentation results of the proposed SGGMM based colour only compared to the results of GMM-Stauffer, KDE and Decolor techniques. These results illustrate clearly that SGGMM modelling (with colour only) achieves higher accuracy in favourable change in illumination scenario. Indeed, most parts of the moving objects are successfully segmented to the foreground mask when using the SGGMM model (colour only). In some cases, small parts of the objects are not detected because of their close colour similarity with the background. Some noise is present but can be removed using median filtering. We notice that some blobs with small areas linked with false foregrounds are still detected. These are easily removed using simple size filtering.



FIGURE 3.6: Comparison between binary objects mask of each method for Office sequence



FIGURE 3.7: Comparison between binary masks of each method for the PETS2006 sequence

3.6.4 Segmentation using SGGMM with colour and pixel uncertainties

Figures 3.8 and 3.9 show test samples of the segmentation results when the colour based SGGMM is augmented by pixel uncertainties and used in the segmentation module for dynamic (shaking) background datasets. Using the SGGMM with colour only leads to classifying some of the shaking background objects as foreground. These figures show that objects within the scenes are well segmented to the foreground when applying pixel uncertainty estimation. Indeed, robustness against background variation is much better compared to GMM-Stauffer, KDE and Decolor.

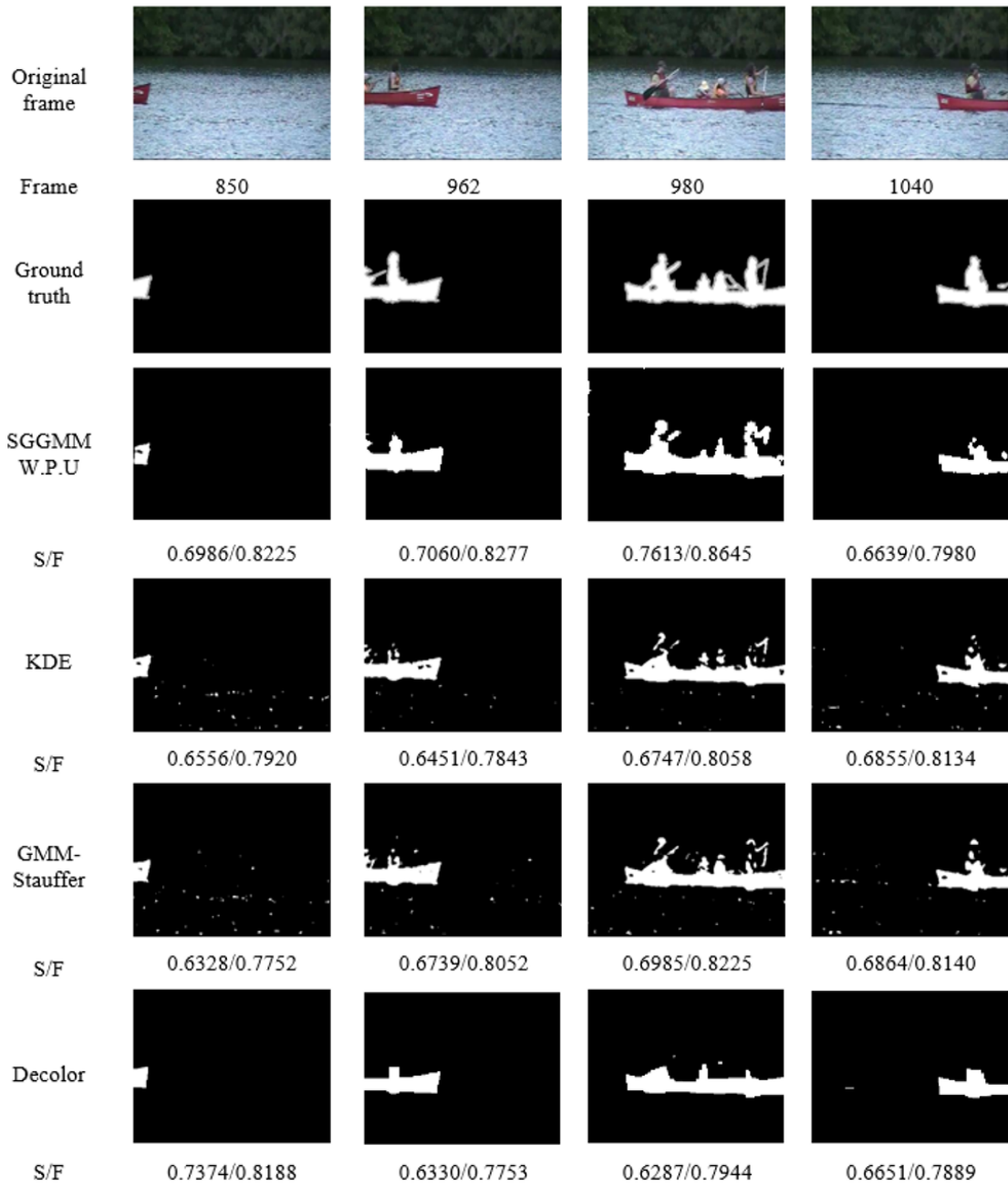


FIGURE 3.8: Comparison between binary masks of each method for Canoe sequence

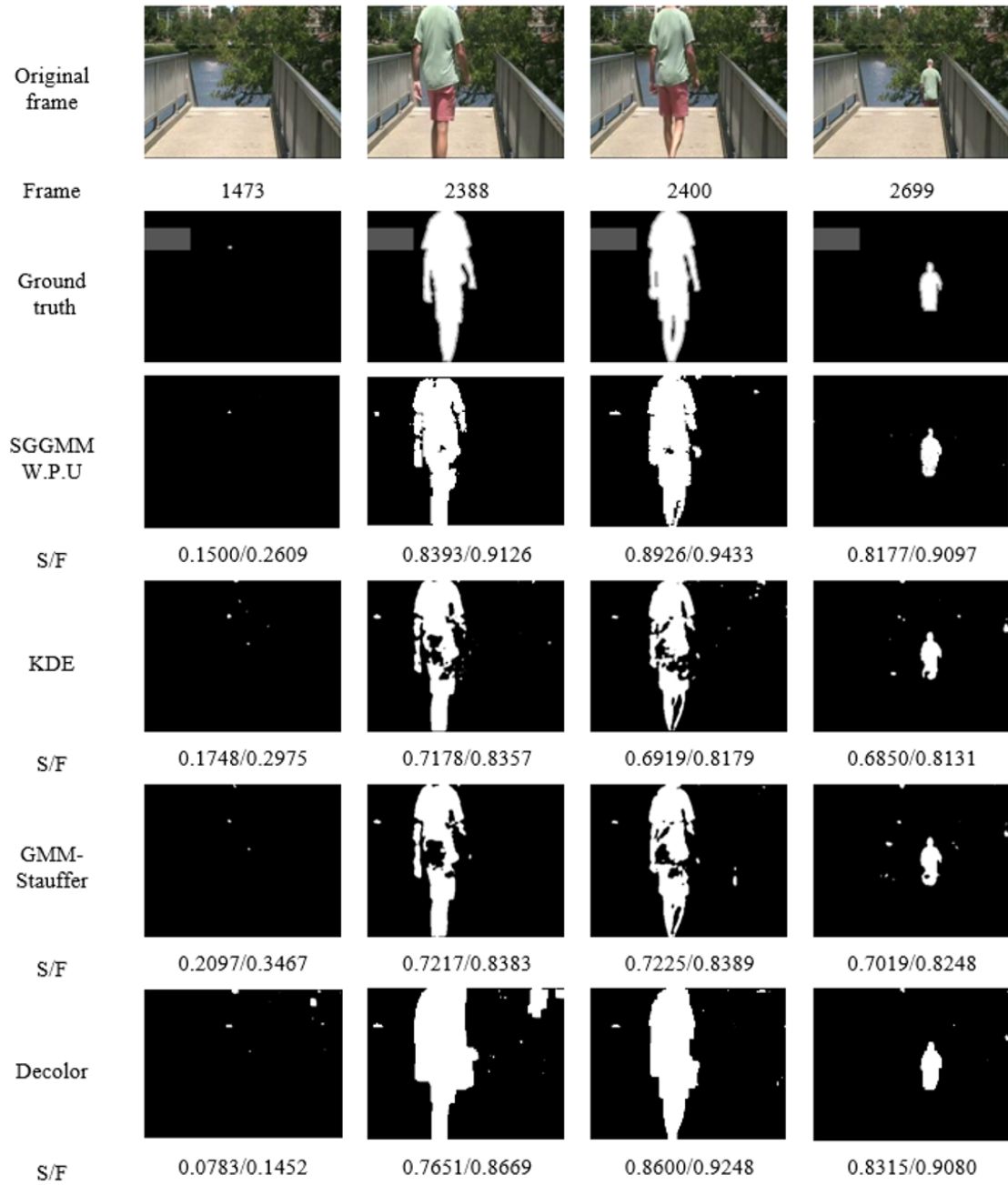


FIGURE 3.9: Comparison between binary mask of each method for the Overpass sequence

3.6.5 Computation performance and real time sensor node implementation

To evaluate the computational efficiency of the proposed approach, a comparison was conducted regarding the execution times between our technique and the other approaches (listed in 3.1). Tests were carried out using the following hardware: CPU: Intel Core i5-2430M -2.4 GHz, RAM: 8.00GB, Operating System: Windows 7. Figure 3.10 shows the execution time per frame for all approaches. It is clear that the execution time per frame for the SGGMM based on colour only is ranked amongst the lowest. This is in addition to its performance when dealing with static camera scenarios. Augmenting the SGGMM with the uncertainties obviously increases the execution time. However, as shown in the previous sub-sections it leads to results with a higher accuracy especially in challenging conditions as shaking background scenarios. Decolor showed the highest computational costs in this study. Driven by its reduced computational cost, our

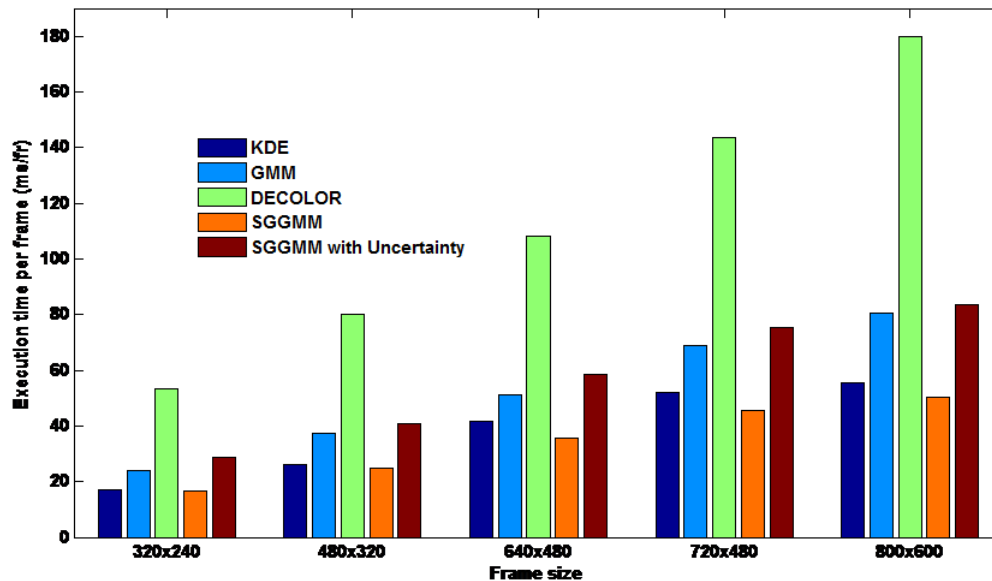


FIGURE 3.10: Execution times of the studied algorithms

method was successfully implemented on the CITRIC camera node [87], (Figures 3.11 and 3.12). The latter consists of a camera board attached to the Telosb

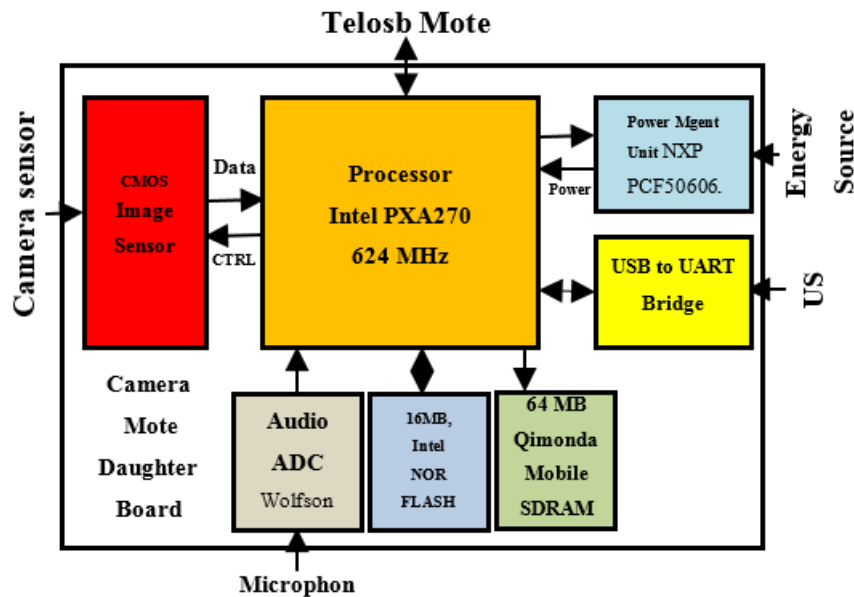


FIGURE 3.11: The embedded camera used in the experiments with its major components

wireless mote. The camera board is composed of a CMOS image sensor, an Intel Xscale PXA270 connected to 64MB of SDRAM, 32MB of NOR FLASH and a power management IC MAX1587. The camera on this platform is the OmniVision OV9655, which is a low voltage SXGA (1.3megapixel) CMOS image sensor. It supports image sizes corresponding to SXGA (1280×1024), VGA, CIF, and any size scaling down from CIF to 4030. The image array is capable of operating

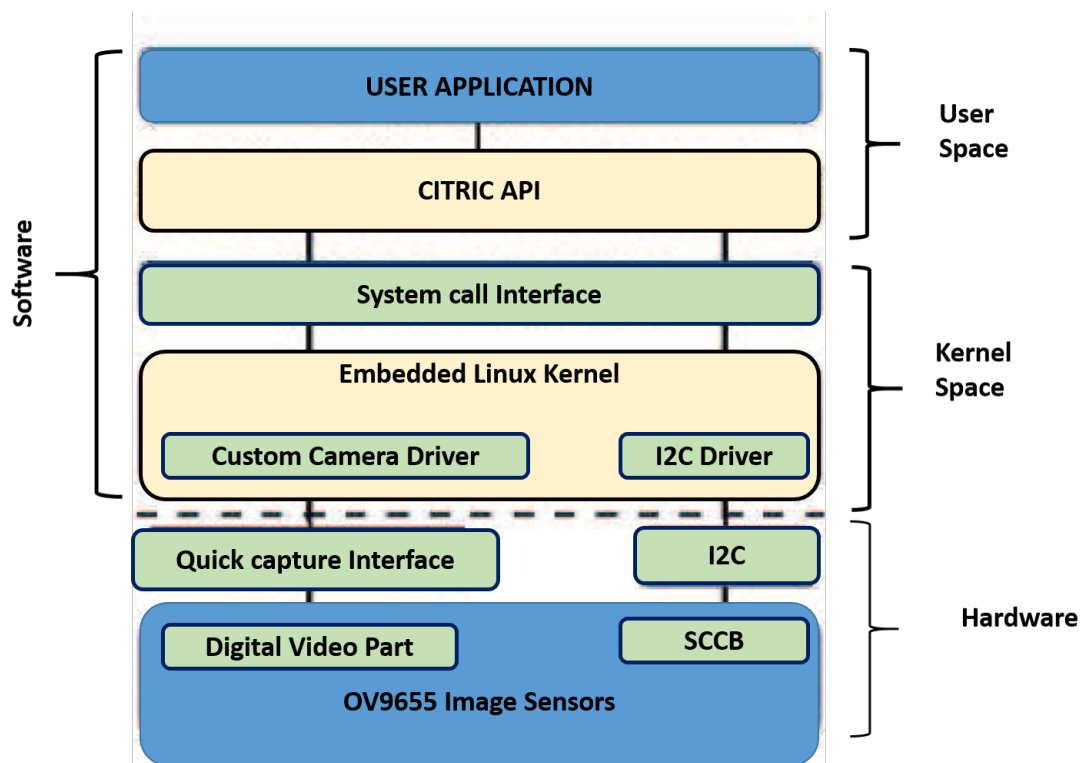


FIGURE 3.12: Software architecture handling the CITRIC camera board

at up to 30 frames per second (fps) in VGA, CIF, lower resolutions, and 15fps in SXGA. The OV9655 is designed to perform well especially in low-light conditions [88].

The camera processor is the PXA270 [89], which is a xedpoint processor with a maximum speed of 624MHz, 256KB of internal SRAM, and a MMX coprocessor to accelerate multimedia operations. The processor is voltage and frequency scalable for low power operation. It can work with a minimum voltage and frequency of 0.85V and 13MHz, respectively. Furthermore, the PXA270 features of the Intel Quick Capture Interface, which eliminates the need for external pre-processors in order to connect the processor to the camera sensor. The CITRIC platform supports variable CPU speeds (208, 312, 416, and 520MHz). The camera has been deployed in several indoor and outdoor environment scenarios, (Figure 3.13). The first scenario was indoors in our Unmanned Autonomous Systems Laboratory

(UASL), Cranfield University, UK. The second and third scenarios were captured outdoor: a public garden and in the street.

	Indoor	Outdoor	
Sequence			
	UASL	Garden	Street
Length	203	85	87

FIGURE 3.13: Number of test frames for the evaluation of the SGGMM using the CITRIC camera

A ground truth at a pixel resolution was required to evaluate the performance of our proposed techniques. Different persons labelled the captured images a number of times and the results were averaged out as described in [85]. Using this ground truth, Similarity (S) and F-measure (F) metrics were estimated and used to appreciate the accuracy in some representative frames. Indeed, Figure 3.14 shows that using the two metrics, for both variants of the proposed approach, provide the expected accuracy. This is said, SGGMM W.P.U obtained better results. Figures 3.15, 3.16 and 3.17 illustrate the performance of the SGGMM colour based model and the superiority of the SGGMM with pixel uncertainties. The F-measure was chosen for the qualitative evaluation as it is the most used to provide a global evaluation of the accuracy of a detection approach.

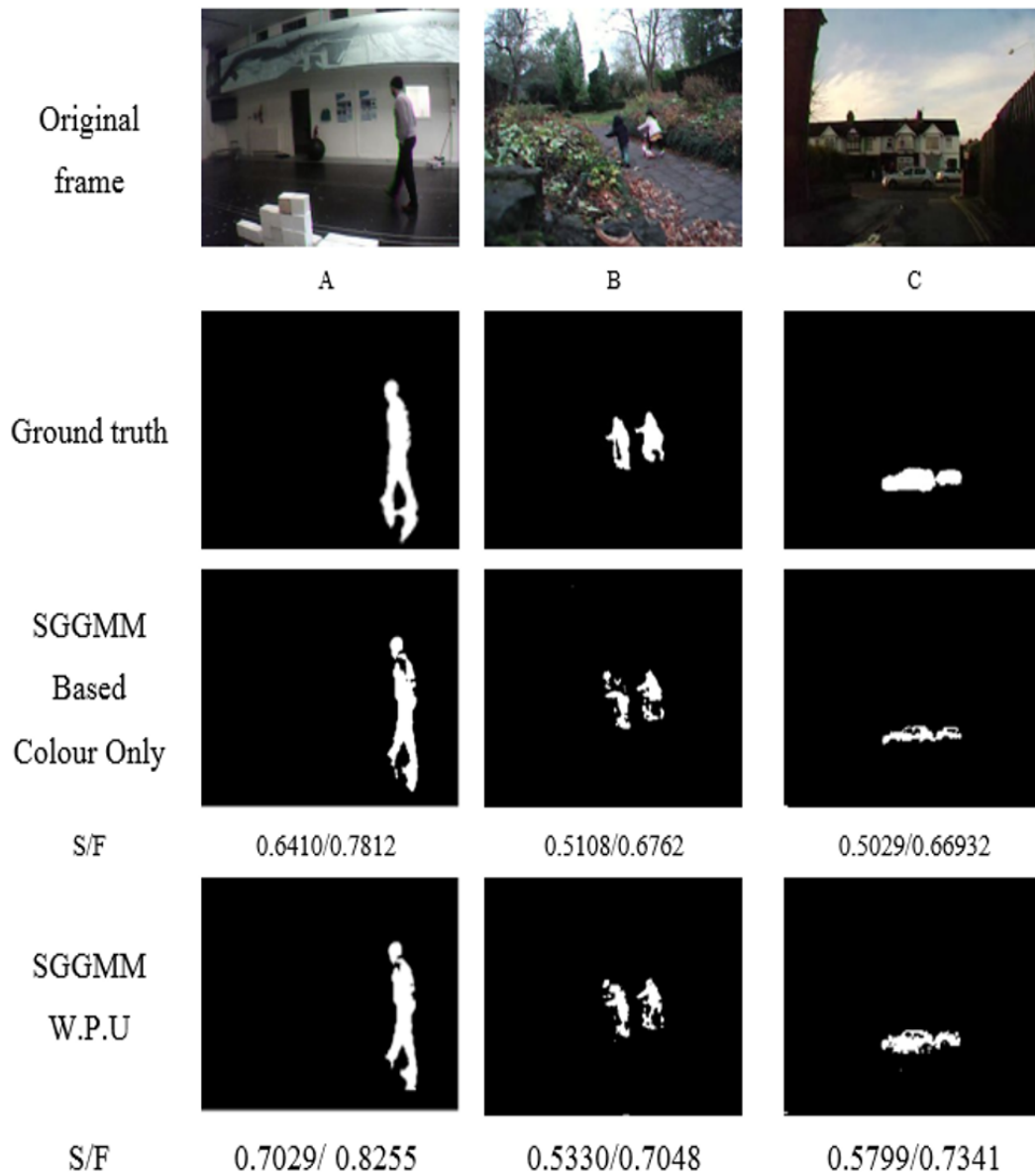


FIGURE 3.14: Comparison between binary masks of the SGGMM based colour only and SGGMM (W.P.U)

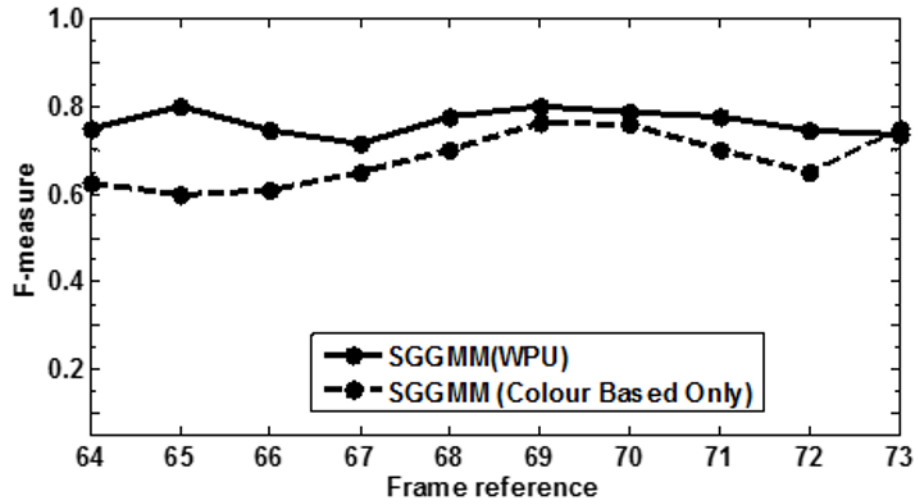


FIGURE 3.15: Evaluation results of the SGGMM based model using for UASL sequence

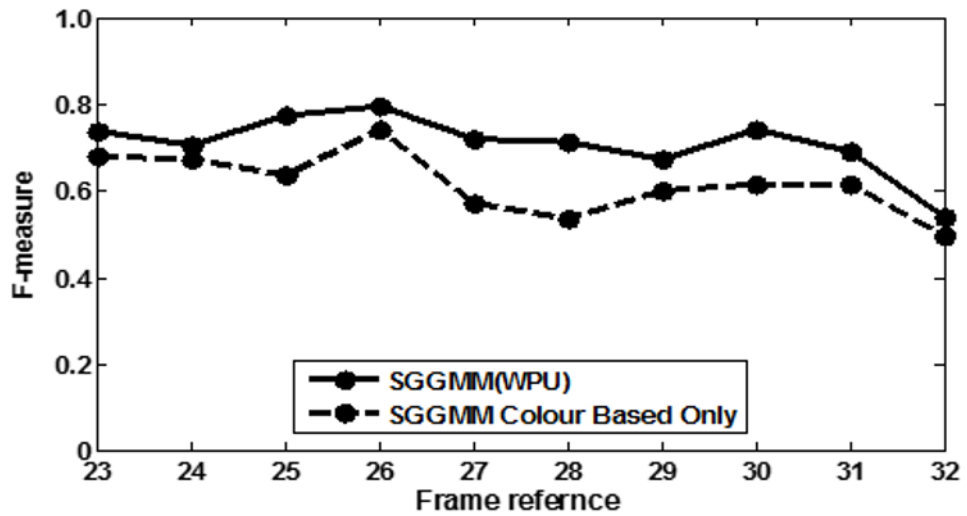


FIGURE 3.16: Evaluation results of the SGGMM based model using for Garden sequence

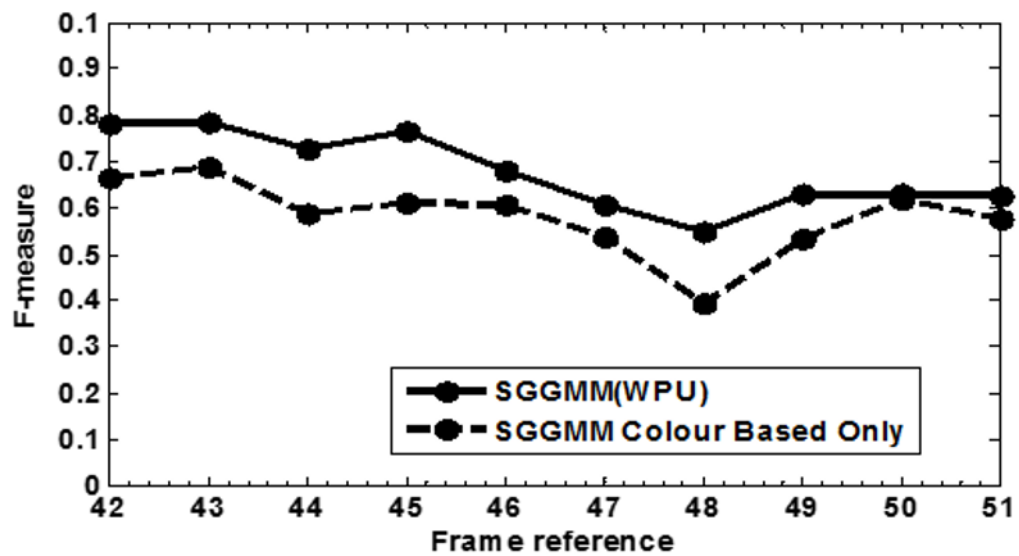


FIGURE 3.17: Evaluation results of the SGGMM based model using for Street sequence

3.7 Conclusion

The problem of visual detection is investigated in this chapter by introducing a new spatially global Gaussian mixture model to approximate the background based on RGB colour. The proposed method adopts a background/foreground subtraction scheme to detect moving objects in image sequences. The SGGMM model is updated by each image of the sequence to take into account scene changes. For improved segmentation performance, pixel location uncertainties have been used to deal with background motion within the scenes. The combination of pixel uncertainties with colours in the SGGMM model resulted in a better overall performance for object detection.

The evaluation of this approach demonstrated the accuracy in detection and the suitability of its implementation in embedded camera sensor network nodes, which includes reduced computation capabilities. The embedded SGGMM colour only model presented comparable performance to the SGGMM W.P.U for static backgrounds. A significant improvement was obtained when combining pixel uncertainties with RGB colour in the SGGMM model for dynamic backgrounds. The latter technique was favourably compared to other segmentation techniques found in the literature.

This page is intendedly left blank

Chapter 4

Moving Object Detection from a Moving platform

4.1 Introduction

In this Chapter, we investigate the problem of motion detection by a moving camera system. Unlike detection using static cameras, and despite the considerable efforts made to investigate this problem, only few proposals are reported in the literature to efficiently address the challenging detection task of detecting in such scenarios. Thus, in this work, we present an approach based on affine image warping using a robust homography method. We further estimate the optical flow after which a Spatial Gaussian Mixture model is used to detect moving objects. The proposed technique combines the efficiency of optical flow for motion detection and the rapidity of execution using the estimated optical flow in a motion compensation-based scheme. This Chapter is organised as follows: A literature review is given in section 4.2. In section 4.3, we present the structure of the proposed solution. In section 4.4, we present the method used to estimate the homography matrix. The technique of optical flow to estimate the velocity

variation in pixel intensities and the adopted approach to cluster these velocities is presented in section 4.5. The final section (4.6) is reserved for the evaluation of our approach using different scenarios for detection from aerial platforms. A summary of the work is given at the end of the chapter.

4.2 Related work

The general approach to the problem of moving object detection in Automatic Video Surveillance Systems (AVSS) is based on the assumption that the camera is stationary (as shown in Chapter 4), so that all frames are registered in the same coordinate system. Consequently, the detection of foreground objects in this situation is performed by background modelling at the first stage, before performing an image subtraction for the second stage [9, 48, 53, 54, 90]. Most of the techniques based on this approach work well with for reasonable change of illumination conditions. Nevertheless, when the camera is moving, the problem becomes more complicated as it is impractical to have a unique background model.

Therefore, in such scenarios, most of the proposed solutions are based on a classical motion compensation technique combined with a background/foreground segmentation technique. The latter, works by comparing the actual captured image with the transformed one and infers differences between the two images for background removal. In the same context, as a second class of solution, the optical flow [91–93] based method are reported to show some efficiency with the complexity of this problem. However, these feature-based methods are used in a way that is computationally costly.

For the class of solution based on a background/foreground segmentation, an approach is presented in [94], in which the author used a background registration technique to construct a background image from the accumulated frame difference information. This step is issued by a separation of the moving object region from

the background by comparing the current frame with the constructed background image. The method is ended by a post-processing step that is applied to the obtained object mask in order to remove noise regions and smooth the object boundary. This method showed usefulness when used for indoor applications. However, no figures are provided to test the method outdoor.

Another approach was reported in [95] where the author proposed a solution based on KLT method for feature detection to estimate the Homography matrix. This was complemented by a single spatio-temporal distribution Gaussian model for motion detection. His solution for background removal was shown to work well in cases of small errors of registration only. In [96], the author suggests the construction of foreground and background appearance models in each frame. Then the posterior of appearance is estimated by computing the product of the image likelihood in the current frame and the prior appearance propagated from the previous frame. Although this technique performed well enough, a primary limitation of the solution was its high computational budget due to the calculation cost of the non-parametric Belief Propagation (BP) used.

Limited research solution using the optical flow for a similar problem are reported in the literature. In these works, the problem of motion detection is approached by assuming that the motion of backgrounds and foreground objects are divided by different optical flows. This approach is adopted in [91], where the Focus of Expansion (FOE) and its residual map for the object of interest detection in the scene is investigated. In [92], the author suggested the use of dense optical flow for detecting moving objects by comparing them to the estimated camera motion. This is achieved by computing the differences between camera motion compensated by backward and forward frames. The frames of interest are then tested against the estimated background models to detect change in pixel intensities. Although this solution showed acceptable performance, the need for processing three (3) successive frames using the dense optical flow required significant computation. Furthermore, as the camera motion should be kept at a low scale, it

makes it difficult for detecting moving objects in real time applications. The problem of motion detection using moving camera was also approached using solutions based on a stereo system, as in [97]. In the latter, the author predicted the depth image for the current time by using ego-motion information and the depth image was obtained from the previous time. Moving objects were detected by comparing the predicted depth image with the one obtained at the current time.

4.3 Structure of the proposed solution

The framework of the proposed solution is illustrated in Figure 4.1. It includes the following three main steps:

- The first stage concerns an affine image warping. This is achieved through:
 - Feature detection and matching operation between two successive images. In this problem, the Speeded Up Robust Features (SURF) detector [98] is used for feature detection.
 - Uncertainties estimation of the detected features using grey pixel intensities. For increased accuracy in the homography, we consider the RGB colours by combining their related uncertainties using the covariance intersection (CI) scheme as proposed in [99] [100];
 - The last task of this stage involves computation of homography matrix using the H_∞ filter [101], as this tool is shown to be efficient in handling problems of estimation with uncertainties in the measurements. [102].
- The second stage considers using a differential local based method for optical flow estimation. This method is the Lukas-Kanade [19]. This choice

is justified by the overall merit of this approach with regard to the globally based approaches in matters of susceptibility to noise, reliability and robustness, in addition to computational efficiency [103].

- In the third step, we compute the optical flow, these are modelled using the spatial Gaussian Mixture Model (SGMM) [104]. The introduction of the SGMM in this context is to determine regions of moving objects. This parametric probability density function represents the estimated optical flow as a weighted sum of Gaussian densities, where the component of greater weight is the one that represents the moving background. Areas that are represented by components of reduced weights represent either the newly introduced pixels or belong to regions of moving objects.

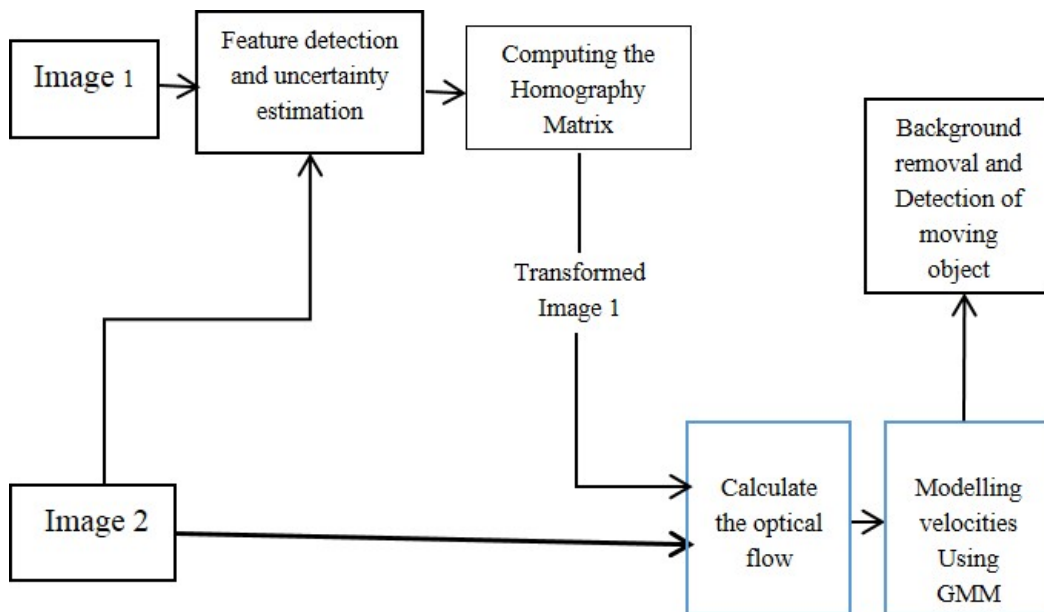


FIGURE 4.1: The overall architecture of the proposed solution

4.4 Robust homography matrix estimation

Computation of the optical flow using the Lukas Kanade method requires a small displacement between pixels of every two successive frames [19] for higher detection accuracy. However, for imaging systems of a moving platform, it is generally impractical to ensure capture sequence of images in the scene with frame rates where displacement between pixels is negligible. To cover this problem, the use of an efficient tool that guarantees higher accuracy in image warping is necessary.

One of the proposed solution to estimate the matrix of homography with high accuracy is to consider the uncertainties in the detected features. This is can achieved by using filtering techniques.

A robust estimator such as the H_∞ filter has demonstrated its ability to deliver better outputs under uncertainties, as opposed to other estimators. To achieve higher accuracy, this filter can use the uncertainties obtained using the RGB colours to detect the features that are initially combined using a fusion method such as the covariance intersection [100].

4.4.1 Feature detection with uncertainty

Feature points are specific and stable points in an image and are extracted using a mathematical operator. Once extracted, these are described in a distinctive way. One of the most prominent detector is the SURF feature detector [98]. It is inspired by the scale-invariant feature transform (SIFT) descriptor [105] while its standard version is several times faster than SIFT and claimed to be more robust against different image transformations than the SIFT [105]. Given a point $X = (x, y)$ in an image I , SURF operates by calculating for each pixel the Hessian matrix $H(X, \sigma)$ [98], that is given by:

$$H(X, \sigma) = \text{Det} \begin{bmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{xy}(X, \sigma) & L_{yy}(X, \sigma) \end{bmatrix} \quad (4.1)$$

Where $L_{xx}(X, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\delta^2}{\delta x^2}g(\sigma)$ with the image I at point X , and similarly for L_{xy} and L_{yy} , σ is taken to be equal to 1.2. Because the Hessian of each pixel is calculated considering the neighbouring pixels, a loss of information about the accurate position of the feature is inevitable. Therefore, a feature location is given with an associated uncertainty. To calculate the related uncertainty, two approaches are commonly used: a residual-based approach or a derivative-based approach [84, 106]. The derivative-based approach is commonly used due to its ease of implementation.

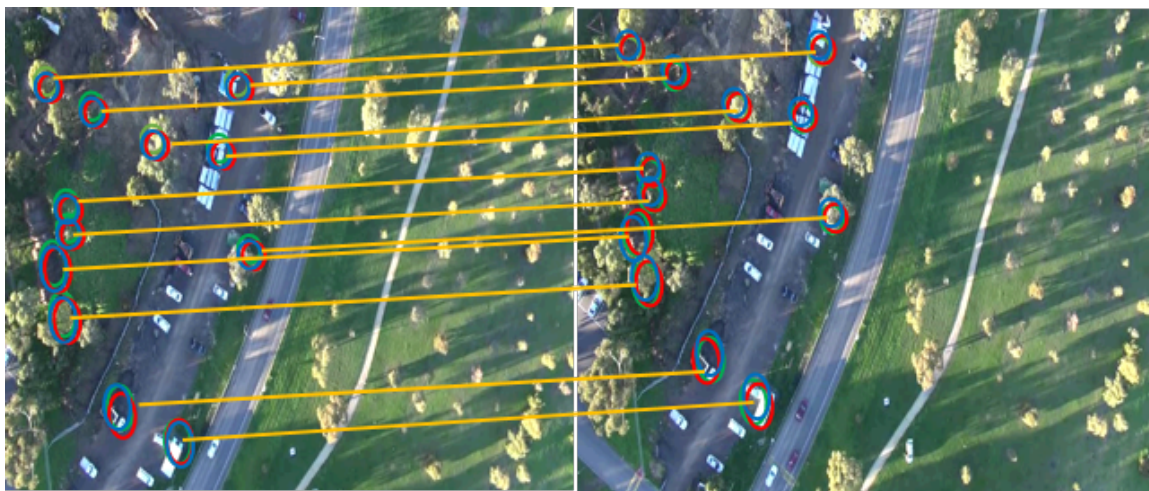


FIGURE 4.2: Matched features and their corresponding Uncertainties from RGB channels¹

The location uncertainty of SURF features using this approach is calculated as the inverse of the Hessian and given by the following:

$$\Sigma = \left(w(i, j)_{i, j \in N_p} \begin{bmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{xy}(X, \sigma) & L_{yy}(X, \sigma) \end{bmatrix}^{-1} \right) \quad (4.2)$$

¹Images source: [107]

with $w(i, j)$ a Gaussian weighting function. To get a more robust estimate it is useful to use an influence region from 3×3 to a 5×5 neighbourhood. Figure 4.2 shows the detected feature with their corresponding error ellipses).

4.4.2 The covariance intersection (CI)

The uncertainty in feature location using SURF detector is estimated from grey images. For more robust uncertainty estimation, feature location errors can be performed over the RGB channels of the captured images. The uncertainties are combined together using the *CI*, which is a fusion rule for combining two or more estimates when the cross correlation between them is unknown. The CI works as follows X_r, X_g, X_b represent the locations of the feature using the RGB channels, with P_r, P_g and P_b as their related covariances. The *CI* fuses these measurements to produce a mean and a covariance pair from the equations:

$$X^{-1} = w_r X_r^{-1} + w_g X_g^{-1} + w_b X_b^{-1} \quad (4.3)$$

$$P = X(w_r X_r^{-1} P_r + w_g X_g^{-1} P_g + w_b X_b^{-1} P_b) \quad (4.4)$$

The resulting estimate is guaranteed to be relevant with $w_i, i = 1, \dots, 3 \in [0, 1]$ with $\sum_{i=1}^3 w_i = 1$. Moreover, it is shown to be optimal for the case where the cross covariance is optimal. The determination of the weighting coefficients is based on an analytical procedure [99].

4.4.3 Modelling the problem of the homography matrix estimation

When the sequence of images is captured by an aerial platform that flies at a high altitude, the assumption that these images are of the same plane is reasonable. Therefore, the relationship between two frames can be described by a homography

matrix H according to the following equation:

$$z' = Hz = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (4.5)$$

where $P = [x, y, z]$ and $P' = [x', y', z']$ are the coordinates of the matching features, $h_{i,j}$ with $i = 1, 2, j = 1, 2$ are the parameters of H that describe the rotation, h_{13} and h_{23} describe the translational parameter, and h_{31} and h_{32} are related to projectivity. By introducing $x'_2 = x'/z'$ and $y'_2 = y'/z'$, we obtain the following:

$$x'_2 = \frac{h_{11}x + h_{12}y + h_{13}z}{h_{31}x + h_{32}y + h_{33}z} \quad (4.6)$$

$$y'_2 = \frac{h_{21}x + h_{22}y + h_{23}z}{h_{31}x + h_{32}y + h_{33}z} \quad (4.7)$$

By setting $z' = 1$ and rearranging we get:

$$x'_2(h_{31}x + h_{32}y + h_{33}z) = h_{11}x + h_{12}y + h_{13}z \quad (4.8)$$

$$y'_2(h_{31}x + h_{32}y + h_{33}z) = h_{21}x + h_{22}y + h_{23}z \quad (4.9)$$

By putting $h_{33} = 1$ (the scale parameter) and reformulating equations (4.8) and (4.9), we obtain [108]:

$$\begin{bmatrix} x & y & 1 & O & -xx' & -yx' \\ O & x & y & 1 & -xy' & -yy' \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} \quad (4.10)$$

where O is $[0 \ 0 \ 0]$. By changing the notation and making the necessary rearrangement, equation (4.10) can be rewritten as:

$$Y = CX \quad (4.11)$$

where X is the vector of the parameters of the homography matrix H that we aim to estimate. In order to determine an optimal (H), we use a technique of state estimation that takes into account the presence of model uncertainties, as encountered here. This technique is the robust H_∞ filter.

4.4.4 The H_∞ filter

After many years of experience using the kalman for highly reliable systems such as used for spatial navigation, there was a need for a new filtering scheme that can be used for system with measurements of high uncertainties. The new adopted scheme handles modelling errors and noise of non Gaussian type while it minimises

the worst-case estimation errors rather than the covariance of the estimation error. State estimators that can tolerate such uncertainties are called robust.

Although robust estimators based on kalman filter theory can be designed, these approaches are somewhat "Ad-hoc" in that they attempt to modify an already existing technique. In contrast, the H_∞ filter was specifically designed for optimality and robustness. The H_∞ filter as a recursive estimator have received considerable attention in literature due to its wide range of applications.

Unlinke the classical estimator such as Kalman filter which requires an accurate system model and noise statistics, the H_∞ filter does not requires prior knowledge of the noise statistics but finite bounded energies. Additionally, the H_∞ filter tries to minimise the effect of the worst possible disturbances on the estimation errors and therefore it is more robust against model uncertainty. Its relevant equations adapted to our problem are the following [109]:

$$X_{k+1} = X_k + w_k \quad (4.12)$$

$$Y_k = HX_k + v_k \quad (4.13)$$

where w_k and v_k are noise terms with an unknown distribution law with covariances Q_k and R_k , respectively. We attempt, as designed by the filter, to design a state estimator of the form:

$$Z_k = LX_k \quad (4.14)$$

where L is a user defined matrix (assumed to be full rank), as want to directly estimate X_k then we set $L = I$. The estimate \hat{Z}_k is found after minimising the cost function J as $J < 1/\theta$ where θ is performance bound defined as

$$J = \frac{\sum_{k=0}^{N-1} \|Z_k - \hat{Z}_k\|^2 S_k}{\|X_0 - \hat{X}_0\|_{P_0^{-1}}^2 + \sum_{k=0}^{N-1} (\|w_k\|_{Q_k^{-1}}^2 - \|v_k\|_{R_k^{-1}}^2)} \quad (4.15)$$

where P_0 , Q_k , R_k and S_k are chosen matrices with the condition of being symmetric, positive definite. Hence, the min-max problem is finally defined as:

$$j^* = \min_{\hat{Z}_k} \max_{w_k, v_k, x_0} J \quad (4.16)$$

The worst-case is obtained when w_k , v_k and x_0 are chosen to maximise J . The solution then is to find an estimate \hat{Z}_k which minimises this maximum. This leads to the filter description below:

$$\bar{S}_k = L_k^T S_k L_k \quad (4.17)$$

$$K_k = P_k [I - \theta S_k P_k + H_k^T R_k^{-1} H_k P_k]^{-1} H_k^T R_k^{-1} \quad (4.18)$$

$$\hat{X}_{k+1} = \hat{X}_k + K_k (y_k - H_k \hat{X}_{k-1}) \quad (4.19)$$

$$P_{k+1} = P [I - \theta S_k P_k + H_k^T R_k^{-1} H_k P_k]^{-1} + Q_k \quad (4.20)$$

Note that the output and the input are Y and C in Equation 4.11.

To evaluate the accuracy of the proposed solution, comparison between the standard RANSAC and least square (LS) method, the H_∞ filter with uncertainties from grey intensities, and the H_∞ filter with uncertainties using RGB colours combined with CI is given in Table (4.1). In the latter, the Average Back Projection Errors (ABPE) measure is the criterion of comparison given in (4.21).

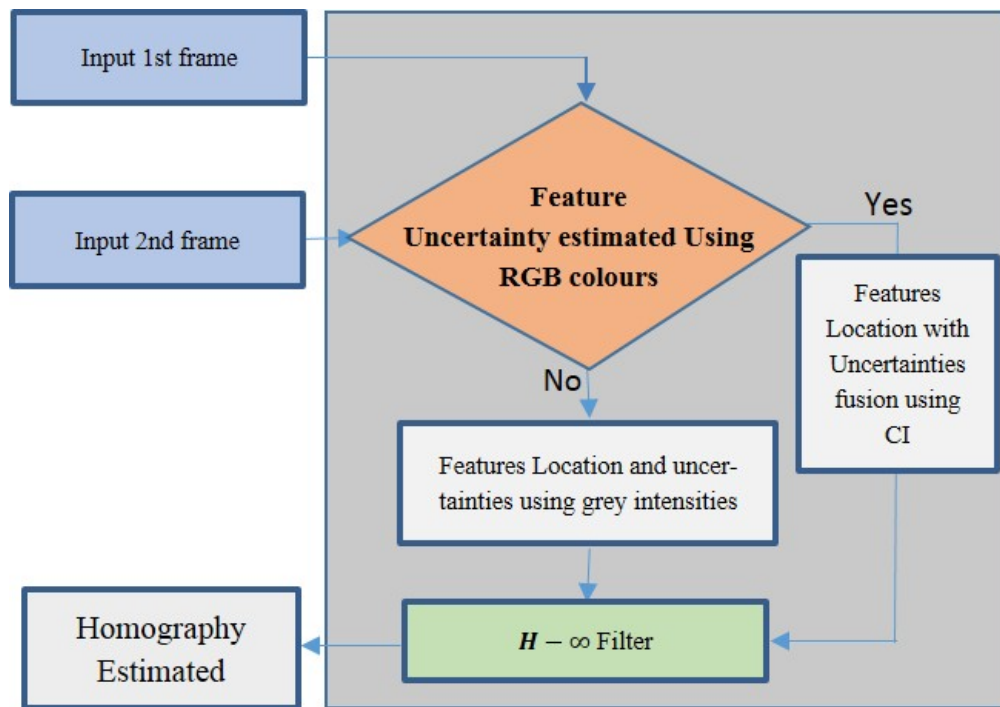
$$ABPE = \frac{\sum_{i=1}^n \|P_i - H^{-1} P'_i\|}{n} \quad (4.21)$$

with P represents the coordinate of features in the previous image, while P' is the coordinate of features in the current image; H is the transformation matrix, while n is the total of number of features. The images used for the tests were obtained from the data set in [85]. Table (4.1) highlights the efficiency of the proposed solution in estimating the Homography matrix. It can be clearly seen that the

H_∞ with uncertainties using RGB colours delivers higher accuracy compared to the H_∞ with uncertainties from grey intensities. We can also notice that both solutions based on the H_∞ gives better results compared to the standard Ransac with least square method. Figure 4.3 refers to the two cases in which the uncertainty in the feature location could be used.

Method of estimating the homography	RANSAC and LS – based method	The H_∞ filter with uncertainties from grey intensities	The H_∞ filter with uncertainties using RGB colours
1 st test	7.1	3.2	2.7
2 nd test	12.9	4.1	2.5
3 rd test	11.3	5.2	3.6

TABLE 4.1: Average back projection error estimation obtained from using different methods

FIGURE 4.3: Estimating the Homography matrix using the robust H_∞

4.5 Motion detection using Optical Flow

The optical flow is defined by the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and the scene [19]. Two main approaches are commonly used for optical flow computation, these are either global or local methods. The former ensures the constancy constraint with a regularising term imposing global smoothness assumptions on the image and/or the flow. The latter uses a differential approach, which consists of finding an optimal solution that minimises an objective function [103].

4.5.1 Estimation of the optical flow using local methods

To minimise an objective function which represents assumptions regarding constancy of selected features in the two images, a local based method works by

considering the image intensity as features. Additionally, the brightness intensity of each pixel is assumed to remain constant even if its position is changing. Therefore, by considering a scalar-valued image sequence $I(x, y, t)$, where (x, y) is the location within a rectangular image domain $\Omega \in R$ and $t \in [0, T]$ denotes time, the reformulation of the brightness constancy assumption can be given as:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \quad (4.22)$$

By using the Taylor series, the expansion of the right-hand side of (4.22) yields:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \frac{\partial I}{\partial x} \delta x + \frac{\partial I}{\partial y} \delta y + \frac{\partial I}{\partial t} \delta t + h.o.t \quad (4.23)$$

where h.o.t stands for higher order terms, which are typically neglected because of their insignificant values due to small motion.

From Equation (4.22) and (4.23), we can write:

$$\frac{\partial I}{\partial x} \delta x + \frac{\partial I}{\partial y} \delta y + \frac{\partial I}{\partial t} \delta t = 0 \quad (4.24)$$

by dividing both terms of (4.24) over δt , the optical flow constraint equation can be written as:

$$I_x u + I_y v + I_t = 0 \quad (4.25)$$

and it can be reformulated as:

$$\nabla I^T \mathbf{u} + I_t = 0 \quad (4.26)$$

where the vectors $\nabla I = [I_x, I_y]^T$, and $\mathbf{u} = [u \ v]$ denote the partial derivatives and the displacement field (also known as the optical flow) respectively.

To estimate the optical flow at local windows R of size $n \times n$, the Equation (4.26) has two unknowns for any given pixel, which means that measurements at the

single-pixel level are under-constrained and cannot be used to deliver a unique solution for the two-dimensional motion at that point. Therefore, a solution is obtained by solving an over-determined set of equations. Assuming that the flow is constant within the neighbourhood of R , the system of equations can be written as:

$$E_{local}(\mathbf{u}) = \sum_R w(p)(E_{Data}) \quad (4.27)$$

where p is the centre of the region and $w(p)$ is a weighting function that gives more influence to the constraints at the centre of the local neighbourhood than those at the border, E_{Data} is the left term of Equation (4.26) and is known as the data term. Taking E_{Data} solely as the regular brightness constraint and assigning equal weights for all pixels in the region, Equation (4.27) can be reformulated so that:

$$\sum_R [\nabla I^T \mathbf{u} + I_t]^2 = 0 \quad (4.28)$$

For a neighbourhood of size $n \times n$, the solution of Equation (4.28) is given as [103]:

$$\mathbf{u} = [A^T A]^{-1} A^T B \quad (4.29)$$

where $A = [\nabla I^1, \dots, \nabla I^{(n \times n)}]^T$ is of size $(n \times n) \times 2$;

while $B = -[I_t^{(1)}, \dots, I_t^{(n \times n)}]^{-1}$ is of size $(n \times n) \times 1$.

One of the well know algorithms that is used to compute the optical flow using this approach, is the Lucas-Kanade algorithm [19]. This method relies only on local information that is derived from a small window surrounding each point of interest. However, one of its disadvantages is that large motions can move points outside of the local window. Therefore, it becomes impossible for the algorithm to make accurate computation. Thanks to the pyramidal LK algorithm [110] which is used to resolve this issue. The Algorithm organises the image as a pyramid of four levels of resolutions and starts processing from the highest level of an image pyramid that contains the lowest resolution and works down to lower levels with the finer details. As a result, scanning over image pyramids allows large motions

to be caught by local windows, and therefore commutation of the pixels velocities [111]. Figure 4.4 shows the obtained results using the optical flow method with the proposed registration technique (H_∞ with Feature uncertainties using RGB colours), with figure (a) and (b) present a sequence of two images with a car moving in the scene. Figure (c) and (d) show the variations of the optical flow components in the scene. It can be seen that variation of the estimated optical flow in the detected regions of motion are not similar. Hence, further processing is needed to detected moving targets with higher accuracy. For this reason, we propose modelling the optical flow components using a clustering scheme adopted in the spatial Gaussian gaussian mixture model (SGGMM) presented in Chapter 3.

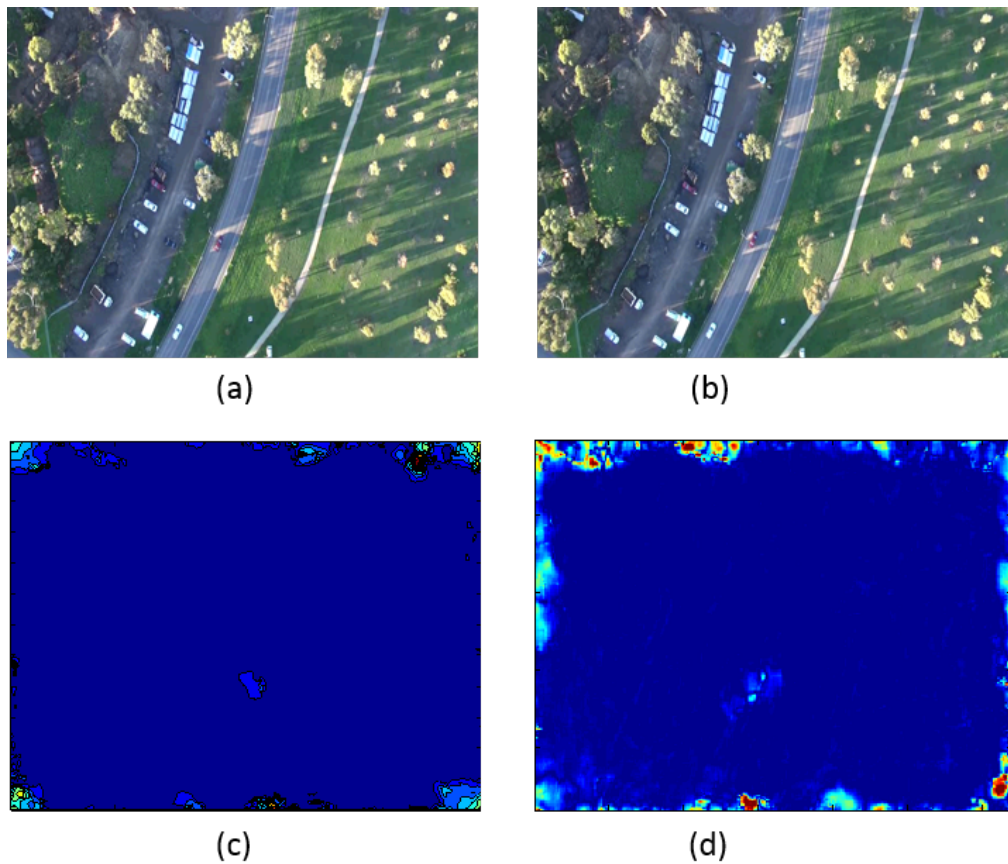


FIGURE 4.4: Optical flow estimation for dynamic background: a) first image, b) second image, c) the u element of the velocity vector, d) the v element of the velocity vector

4.5.2 Modelling optical flow using the SGGMM

In this section we introduce the SGGMM model to determine regions of moving objects from the 'noisy' estimated optical flow. In an optical flow representation, each image pixel value is represented in a feature space by a $2D$ vector $\mathbf{u} = [u \ v]^T$, the values of each component of these vectors are represented by a spatial Global Gaussian mixture model of N Gaussians in 1-dimensional space as follows:

$$p(x) = \sum_{i=1}^N [w_i g_i(x, \mu_i, \sigma_i)] \quad (4.30)$$

where μ_i and σ_i are, respectively the spatial mean vector and variance the i^{th} distribution, and w_i is an estimate of the weight that reflects the likelihood that the corresponding distribution accounts for pixel velocity and satisfies the criterion:

$$\sum_{i=1, n} w_i = 1 \quad (4.31)$$

Each Gaussian distribution $g_i(x, \mu_i, \sigma_i)$ of the mixture is defined as:

$$g_i(x, \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \quad (4.32)$$

where x is represent the component velocity.

The determination of the components of the Gaussian mixture in Equation (4.30) is achieved through the hierarchic clustering scheme presented in Section 3. After modelling the optical flow by the SGGMM, the next step considers definition the pixels belonging to the moving objects in the background. To this end, the cluster of the highest weight is considered as the one presenting the background. The remaining clusters that are not in the contour area are considered as belonging to the moving object. A Final step is to combine the output of the spatial Gaussian segmentation of the two optical flow components (u and v) to eliminated unwanted noise and to improve the the accuracy of detection. Figure 4.5.2 depicts the

different steps processed to detect moving object using the velocity components of the optical flow.

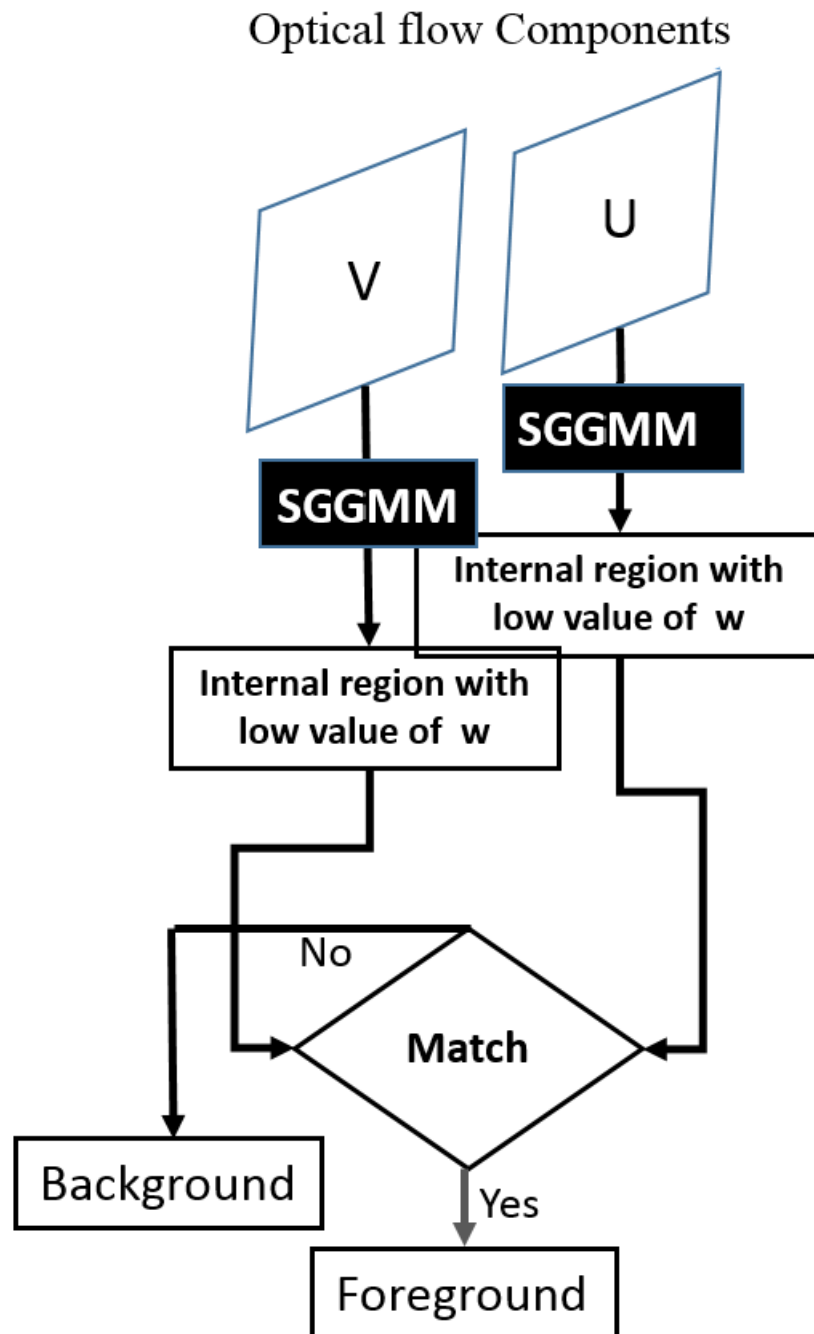


FIGURE 4.5: Proposed scheme for moving object detection

4.6 Experimental results

To show the effectiveness of the proposed algorithm, several tests were performed on sequences obtained from well-known data sets acquired using Unmanned Aerial Vehicle (UAV) cameras. An i5 2.3 GHz laptop computer is used to implement the proposed method by processing 320×240 resolution images. The performance of our method (homography with RGB colours uncertainty and optical flow-HRGBO) is compared to the single Gaussian method (SG) [112]. The dataset used to test the performance of the UAV scenario is obtained from [113]. A quantitative evaluation of the proposed method is given in Table 4.6. Having evaluated both methods for 20 frames, the obtained results clearly shows an overall improvement in the accuracy of detection of the proposed method, especially for the test in the second dataset. Qualitative evaluation of the proposed method compared

		SG	HRGBO
Data set 1	S	0.2601	0.3104
	F	0.4102	0.4804
Data set 2	S	0.1104	0.2951
	F	0.1521	0.3823

TABLE 4.2: Qualitative evaluation of the proposed scheme

to the Single Gaussian is shown in figures 4.6, 4.6, 4.8, 4.9. we can clearly note the improvement in the detection accuracy of the HRGBO compared to the single Gaussian method. The metrics used for the evaluation here are the similarity (S) and the F-measure (F) described in Section 3.

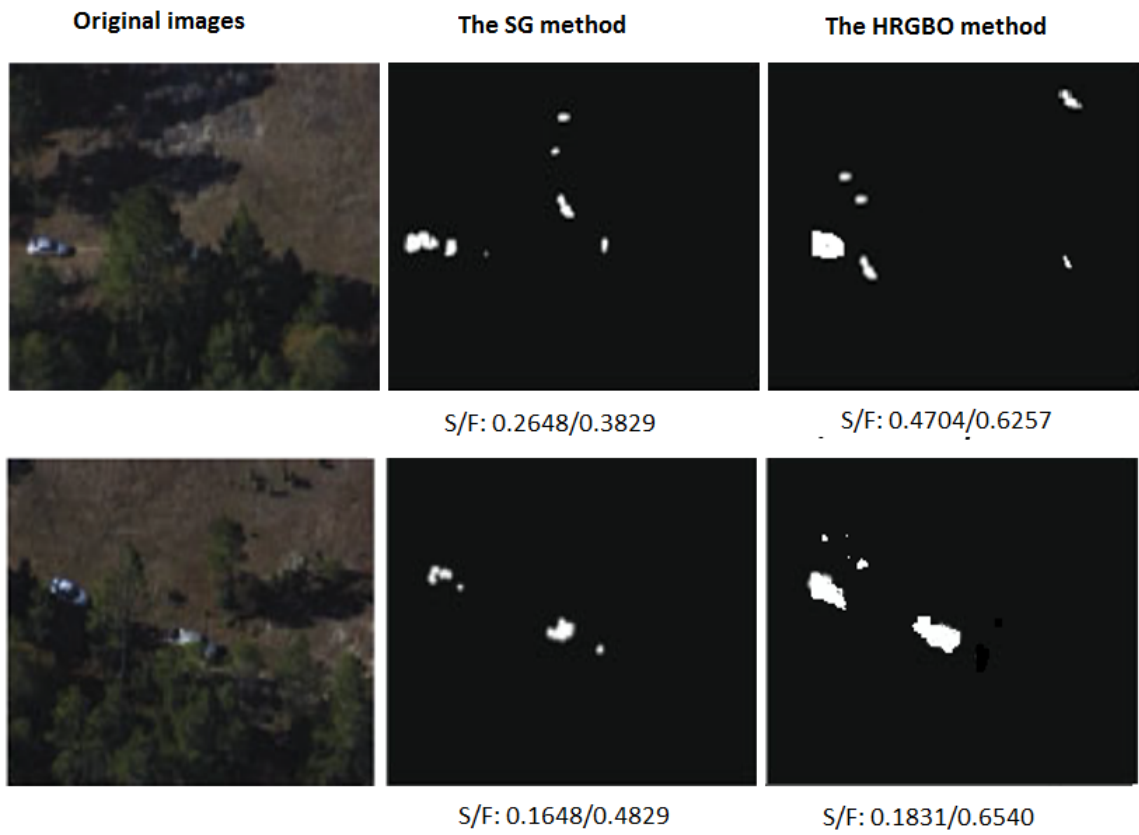


FIGURE 4.6: Qualitative evaluation of the proposed method-first scenario, first test²

²S/F: Similarity/F-measure

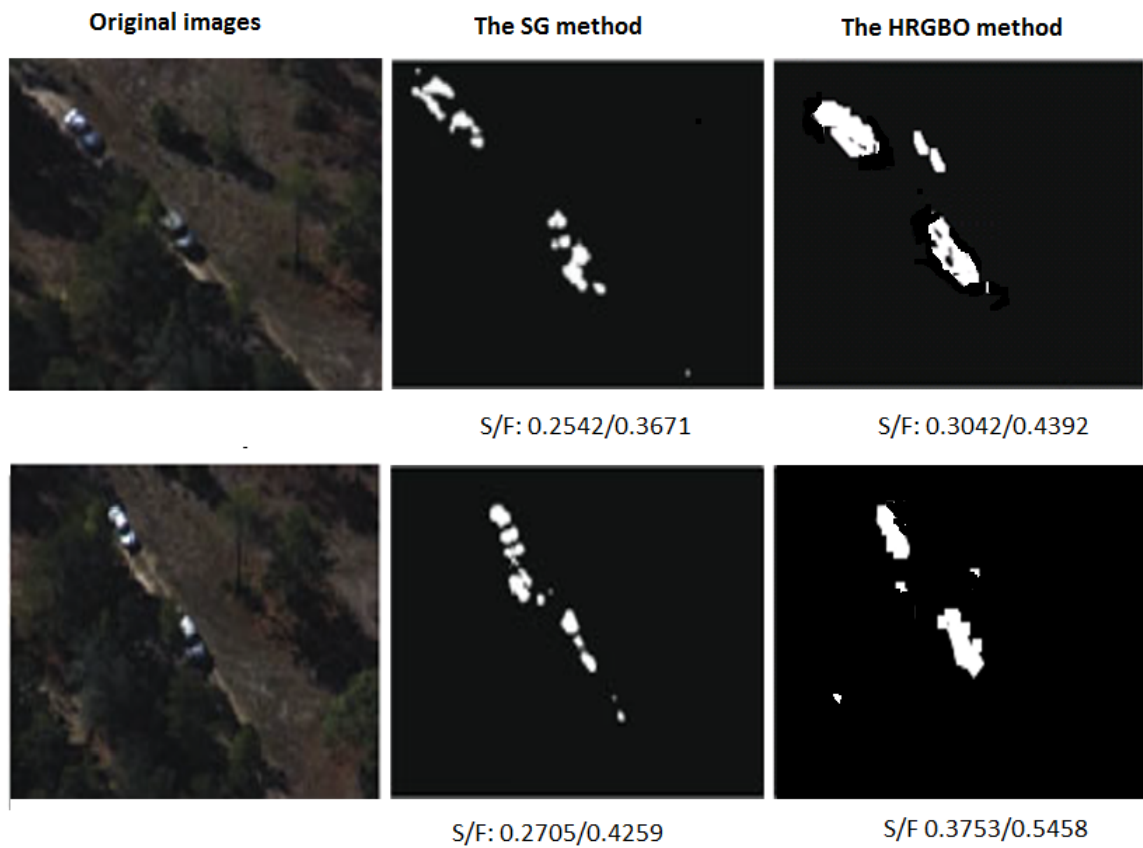


FIGURE 4.7: Qualitative evaluation of the proposed method-first scenario, second test

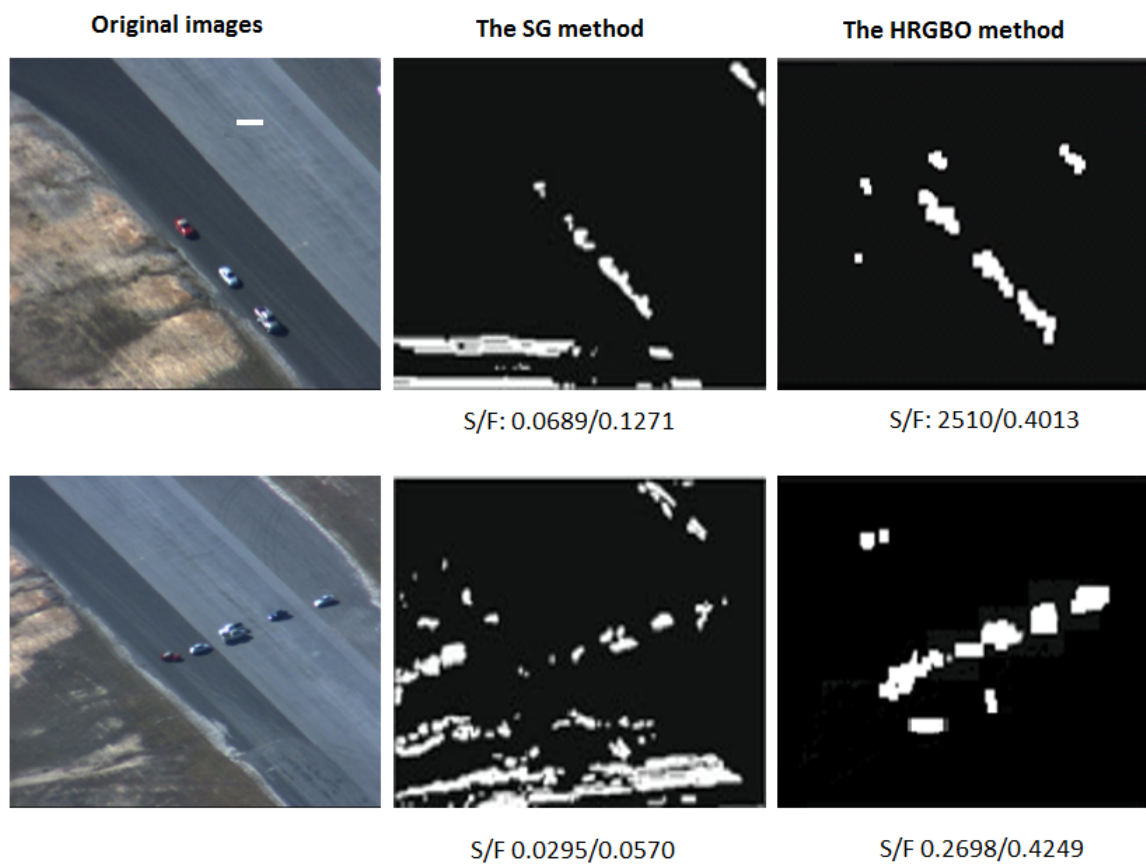


FIGURE 4.8: Qualitative evaluation of the proposed method-second scenario, first test

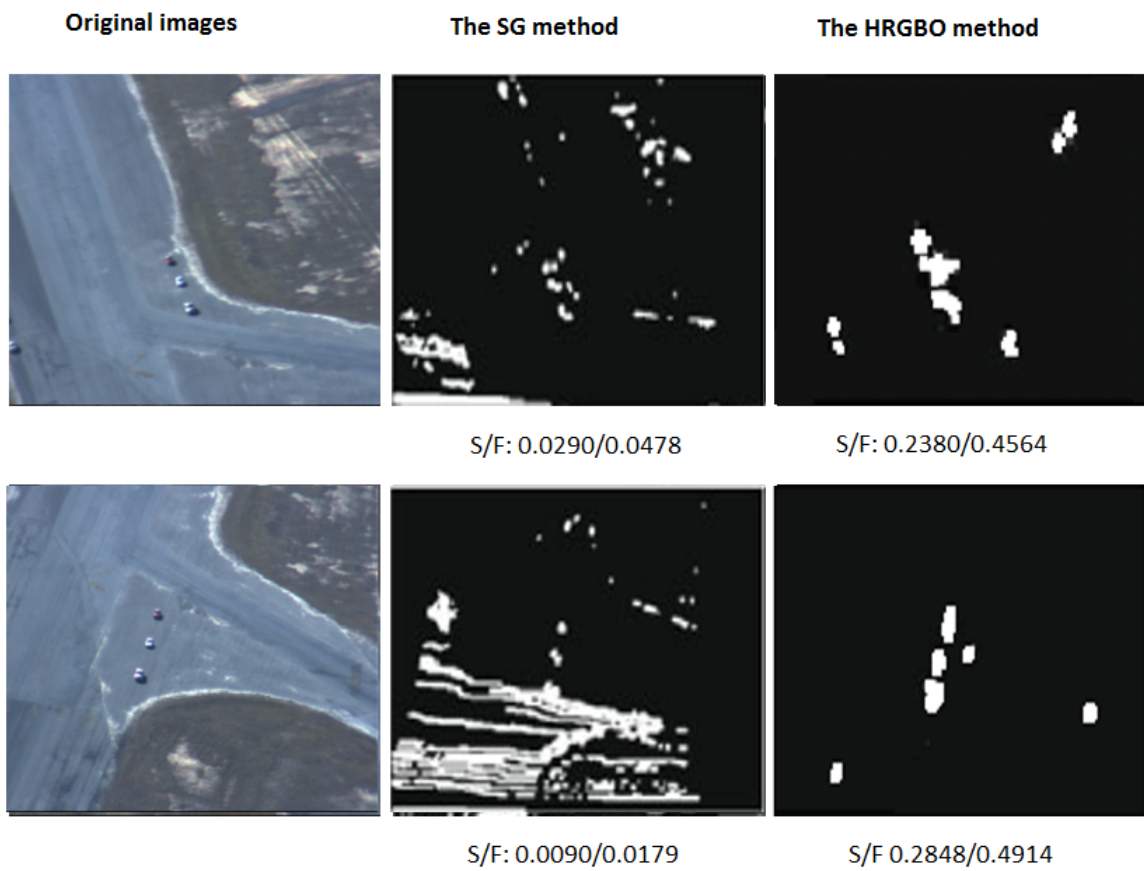


FIGURE 4.9: Qualitative evaluation of the proposed method-second scenario, second test

4.7 Conclusion

In this Chapter, we presented a novel approach to detect moving objects from moving cameras. We have shown that the optical flow method can be used to solve the problem of motion detection from moving platforms, when used in motion compensation-based scheme.

To ensure that the underlying intensity constancy assumption of the optical flow, we proposed the use of the robust H_∞ filter with feature uncertainties for image registration.

Despite the efficient tools used to counter miss-registration issues, the overall estimated optical flow is shown to be corrupted with noise that vary from frame to frame. To address this problem, we proposed the use of the SGGMM to improve the accuracy of detected regions corresponding to moving targets between two successive frames.

The effectiveness of the proposed solution has been demonstrated using sequences that represent two different scenarios. These consider detection of moving cars using an aerial surveillance system.

This page is intendedly left blank

Chapter 5

Robust Acoustic Source Localisation in Low Cost Sensor Networks

5.1 Introduction

The two previous chapters were dedicated to visual change detection with consideration of some particular issues in embedded vision systems. As a main focus, in this thesis, is to investigate the development of novel techniques of fusion between the visual and the acoustic data, we will investigate, in this chapter, the problem of the acoustic event localisation in a different type of embedded systems, which is the wireless sensors networks. More specifically, we will investigate the problem of development of an efficient and robust algorithm for acoustic source localisation based on the Time Delay Of Arrival (TDOA) estimation in the context of low-cost Wireless Sensor Networks (WSN). Part of the available solutions in the literature, for such a challenge formulate this problem as a minimisation of a non-linear least square function, which is solved using the Gauss-Newton method. The

latter shows a degraded performance especially when it is initialised far away from the desired solution. To make up for this inefficiency, we propose to adapt for this minimisation a trust region based optimiser named Powell's Double Dogleg under a Total Least Squares (TLS) framework. Furthermore, we characterize, for the first time in the literature, the uncertainties available in the TDOA measurements and propose a new way of evaluating them experimentally. These uncertainties are taken formally into account in the proposed optimizer through the adoption of weighted norms in its optimisation process. Evaluation results based on a source localisation setup demonstrate the suitability of the proposed algorithm in terms of the overall accuracy and the global convergence rate.

5.2 Related works

Acoustic sources localisation in Wireless Sensor Networks has been well a investigated subject in recent years. This is due to its potential applications ranging from the military domain to those that target the civilian sectors. Indeed, research in this developing technology has resulted in very interesting solutions, such as a vehicles localisation system proposed in [114]. The latter exploits the acoustic signature made up by the vehicle's engine noise. Another important application is the gun firing and sniper localisation system investigated in [115], which aims to localise firing sources using the acoustic information. As an example of this technology utilisation in the civilian domain, an efficient system that uses the acoustic information for tracking wild animals in their natural habitat can be found in [116].

The challenge in the task of acoustic localisation comes from the fact that an accurate localisation using TDOA measurements can be achieved only if the acoustic sensors positions are known with high accuracy, with perfect time synchronisation between sensors, while the circular propagation of acoustic waves

should be known and be of constant speed. In a distributed sensors network, however, these assumptions are barely satisfied. Different techniques have been proposed to deal with the proposed problem that is challenged by measurements uncertainties [45, 117–119].

A class of the proposed solutions to this problem formulated it as a minimisation of the sum squares of errors of a linear problem [120]. This linear approximation is very conservative and is based on non-realistic assumptions that would lead to inaccurate results. Another class of solutions was formulated as a minimisation of the sum squares of errors of a non-linear problem. The solution to this problem is generally approached using linear search method, which is the Gauss-Newton method [121]. While this technique is featured by an ease of implementation with fast computation, one of its major drawbacks is that it does not converge to a global minimum unless it is initialised sufficiently close enough to the solution. It is also not considering the effect of the measurement uncertainties in the optimisation process.

In this chapter, we develop a theoretical model that is validated experimentally to approximate the available uncertainties in the source localisation process. To our knowledge this is the first time that such uncertainties have been estimated. Adopting the Total Least Squares (TLS) optimisation framework, as proposed in this paper, provides robustness towards dealing with noisy data. While the name total least squares has appeared only recently in the literature [122–124], this framework of optimisation was known in the past by different names such as orthogonal regression, errors-in-variables, and measurement errors. The univariate problem for this type of optimisation was presented in 1877, [125]. Only thirty years ago, this technique was extended in [126] and in [127] to multivariate cases. Under this framework, we innovate by proposing an alternative optimizer algorithm based on the trust region technique which is the Powell’s Dogleg [128].

While the original version of this iterative method (Powell's Dogleg) combines between steps estimated using the Steepest Decent and the Gauss-Newton method, we improve this technique by investigating the use of weighted norms in calculating these steps. These weights are obtained through evaluating the uncertainties in the TDOA measurements. We compare the performance of the different version of the proposed method to the Gauss-newton method. This comparison involves the level of accuracy, the speed of convergence in addition to its computational cost with increased number of sensor nodes.

Section 5.1 presents the signal propagation model in WSN using a TDOA based approach, while Section 5.4 presents the proposed uncertainty model due to errors of synchronisation in WSN. We present the Powell's Dogleg/Double Dogleg algorithm with the proposed improvement in Section 5.5 with evaluation results in Section 5.6. The overall summary of the work is given in the conclusion.

5.3 TDOA based localisation signal model In WSN

TDOA based acoustic source localisation approach is shown to be more practical than the energy based method [43]. This is mainly because it does not require a prior knowledge of the signal generated by the sound source. The motivation for this approach comes from the observation that the sound wave propagates at constant speed (sound speed) from the acoustic source to the listeners (the acoustic sensors). The mathematical modelling of the TDOA measurements at the different sensor pairs in WSN enables the estimation of the source position. Let us consider a WSN composed of n acoustic sensors that form a number of $(n-1)$ sensor pairs with $r_i = [x_i, y_i, z_i]^T \in R^3$ representing the position of the acoustic sensor i . The aim of the source localisation approach using TDOA measurements is to precisely determine the sound source location $s = [x, y, z]^T$ (where T denotes

the matrix transpose) by utilising $(n - 1)$ TDOA measurements obtained using a minimum of $n = 4$ sensor nodes for 3D localisation (while a minimum of 3 sensors are required for 2D localisation). Additionally, the position of the acoustic sensors $r_i = [x_i, y_i, z_i]^T$ with $i = 1, \dots, n$ should be initially known. The TDOA measurements t_{ij} between signals received by a pair of sensors nodes is given by:

$$t_{ij} = t_j - t_i; i, j \in 1, \dots, n \quad (5.1)$$

where t_i, t_j are the times it takes for the signal transmitted by the source s to arrive at the sensors r_i and r_j respectively:

$$t_i = \frac{\|d_i\|}{c}; t_j = \frac{\|d_j\|}{c} \quad (5.2)$$

with $i, j \in [1, \dots, n], c$ is the propagation speed of the transmitted signal, while d_i, d_j are the range vectors for the sensors r_i, r_j respectively. These can be written as the following:

$$d_i = s - r_i; d_j = s - r_j; i, j \in [1, \dots, n] \quad (5.3)$$

From Equation (5.1) and (5.2), the TDOA t_{ij} , can be written as:

$$t_{ij} = \frac{1}{c} (\|d_j\| - \|d_i\|); i, j \in 1, \dots, n \quad (5.4)$$

Writing the TDOA measurement as a function of possible source locations as given in Equation (5.4) defines elliptic paraboloid (or a hyperbola in 2D localisation (5.1)). The sound source location is obtained from the intersection of three or more elliptic paraboloid defined in the following set of nonlinear equations:

$$f(s) = \begin{cases} t_{12} = \frac{1}{c} \cdot (\|s - r_2\| - \|s - r_1\|), \\ t_{13} = \frac{1}{c} \cdot (\|s - r_3\| - \|s - r_1\|), \\ \vdots \\ t_{1n} = \frac{1}{c} \cdot (\|s - r_n\| - \|s - r_1\|), \end{cases} \quad (5.5)$$

In practice, we deal with noisy measurements \tilde{t}_{1i} . These are defined by:

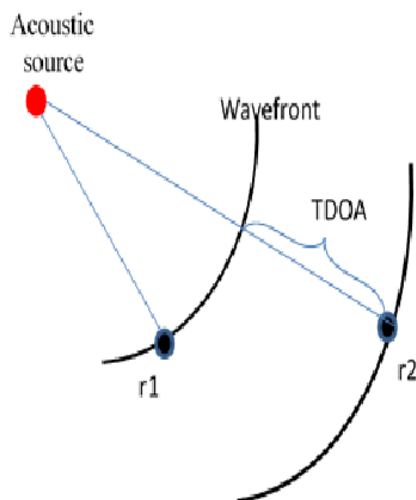


FIGURE 5.1: Two-dimensional TDOA source localisation geometry with two receivers

$$\tilde{t}_{1i} = t_{1i} + \delta t_{1i}; i \in 2, \dots, n \quad (5.6)$$

With δt_{1i} is the TDOA uncertainties in the measured time \tilde{t}_{1i} which are generally assumed to be Gaussians with zero mean. The TDOA noise covariance matrix can be written then as:

$$\Sigma = \begin{bmatrix} \delta t_{12}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \delta t_{1n}^2 \end{bmatrix} \quad (5.7)$$

where δt_{1n} is the TDOA uncertainty in the measurement provided by the sensors pair (r_1, r_n) . Though this assumption might be reasonable in wired systems where the origin of errors is unknown and of a random nature, in a low-cost WSN, the situation is however different. The uncertainty can be shown that it is due to synchronisation issues and the low sampling rate of the sensor nodes that contributes directly to determine the interval in measurements errors of these sensors.

5.4 Uncertainty in TDOA measurements in Low cost WSN

In our WSN based source localisation application a single-hop synchronised WSN using a protocol based on Reference Broadcast Synchronisation (RBS) strategy is applied [39]. The uncertainty in the measured time instant t of the acoustic events by a sensor node i can be given as a form of two independent physical measurements,[129], as following:

$$\delta t_i = \sqrt{\delta t_{sy}^2 + \delta t_{sr}^2} \quad (5.8)$$

With δt_{sy} is the uncertainty in the time reference of the node due to synchronisation issue, while δt_{sr} is the uncertainty due to the limited sampling rate. For TDOA measurements, two basic approaches can be applied: either by relying on the on the detection of the peak energy of the acoustic signal, [115], as we do in this work or using the wavelet transform with envelope signal processing, as proposed in [35].

5.4.1 The Uncertainties due to synchronisation issue

A first factor which contributes to the measurements uncertainties is the timer drift rate (*drift*),[130]. Contrary to the classical notion of clock that offers the opportunity to give a time measurement in relation to a standard reference, the sensors nodes are only fitted with a timer that is able to measure time intervals. These, however, are designed to be crystal based, basically because of their reduced cost. Such crystals are susceptible to large drifts in comparison to the ideal clock. Figure 5.2 shows the increase in the time drift for each of the sensor nodes. As this drift rate varies from one sensor to another, the difference in drift rate between two sensor nodes also increases with time [130]. We are interested in the

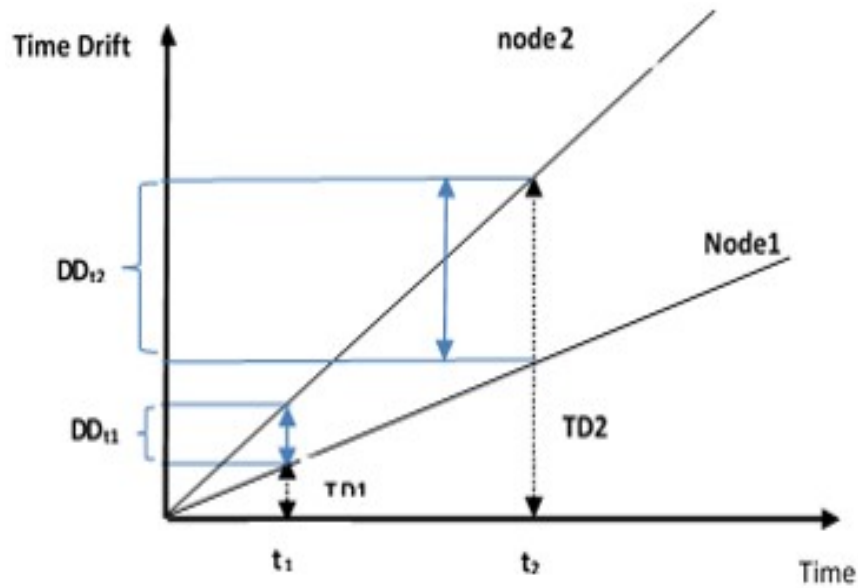


FIGURE 5.2: Drift among the node timers

time difference of measurements between two sensors nodes. Using the reference broadcast synchronisation (RBS) protocol for single-hop network, the scale of errors can be shown to be proportional to the duration between the time of the measurement and the time of the last reference broadcast. By considering the two parameters of synchronisation and variation in the drift rate, the propagation of errors (uncertainty) due to the difference in timers drift rate between two sensors nodes can be formulated as:

$$\delta t_{\Delta drift} = T * \delta drift \quad (5.9)$$

With T is the time interval between references broadcast messages. Evaluation results using simulation shown in Figure 5.3 highlights the proportional relationship between the difference in the drift rate and the length of periods between re-synchronisation events. It can be clearly seen that the average of errors increases relatively due to the progression of time between synchronisation events. In practice, the difference in drift rate $\delta drift$ between the sensors nodes is not the only reason for the errors due to the synchronisation as a time delay in receiving and processing the reference broadcasts messages also may occur. This delay does

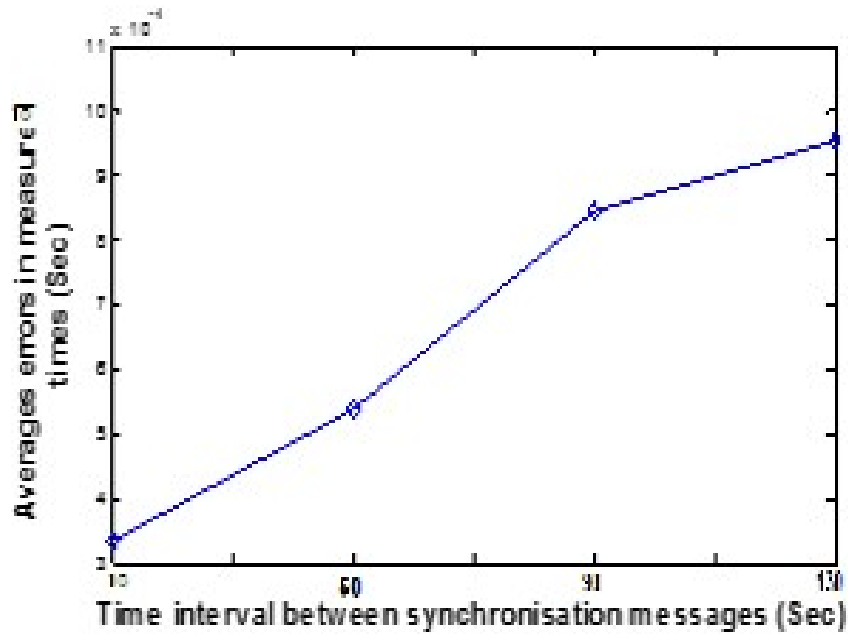


FIGURE 5.3: : Impact of the time interval between synchronisation messages on the uncertainty in TDOA measurements

not occur at the sender level since we are using the RBS synchronisation protocol with the reference broadcasts messages being transmitted to the different sensors nodes at the same time. However, this time delay is available in the reception of the reference broadcasts messages at the sensors nodes level. There are many reasons behind this delay problem such as time of stacking and un-stacking of TinyOS messages. This is in addition to the internal communication issue between the hardware and the application layer. The error due to this time delay is of a random nature and cannot be modelled theoretically by a specific model. Having said that, the error of this delay is taken into account in our study and bounded by a bound δub . The propagation of errors due to the synchronisation issue can be then modelled as the following:

$$\delta t_{sy} = \delta t_{\Delta drift} + \delta ub \quad (5.10)$$

5.4.2 The uncertainty due the low sampling frequency

Since the low cost sensors nodes are designed for computational setups of a reduced budget, the maximum sampling rate we can achieve using sensor nodes of the Mica family such as the MICAZ with the mts310 is 4.8 KHz, [35]. This limitation can be partially made up by using innovative signal processing techniques that work below the Nyquist criterion as proposed in [35] to implement in an advanced acoustic application. However, the uncertainty in time measurement is still significant. Indeed, the error margin, due to this limitation can be written as following:

$$\delta t_f = \alpha \frac{1}{f} \quad (5.11)$$

with f represents the sampling frequency used while $\alpha \in [0, 1[$ determines the uncertainty bounds which are assumed to be uniformly distributed. Results obtained from simulation shown in Figure 5.4 illustrates the disproportional relationship between the average accuracy (errors) in TDOA measurements and the scale of the acoustic signal sampling rate. It clearly shows that accurate TDOA measurements are obtained with higher sampling rate. From the details previously given in sub-sections 5.4.1 and 5.4.2, the uncertainties in TDOA measurements t between two sensors i and j are function of errors due to synchronisation and limited signal processing capabilities (sampling frequency) and can be written as:

$$\delta_{t_{ij}} = \delta_{t_i} - \delta_{t_j} = \sqrt{2(\delta t_{sy}^2 + \delta t_{sr}^2)} \quad (5.12)$$

Thus:

$$\delta_{t_{ij}} = \delta_{t_i} - \delta_{t_j} = \sqrt{2((1/f^2) + (T * \delta drift + \delta ub)^2)} \quad (5.13)$$

An analysis of Equation (5.13) is done using a Monte Carlo simulation, in which a set of 200 measurements have been taken for a pair of sensors nodes at different time instants and with different drift rates $\delta drift$ and time delay δub . This analysis reveals the possibility of modelling these errors using either a form like

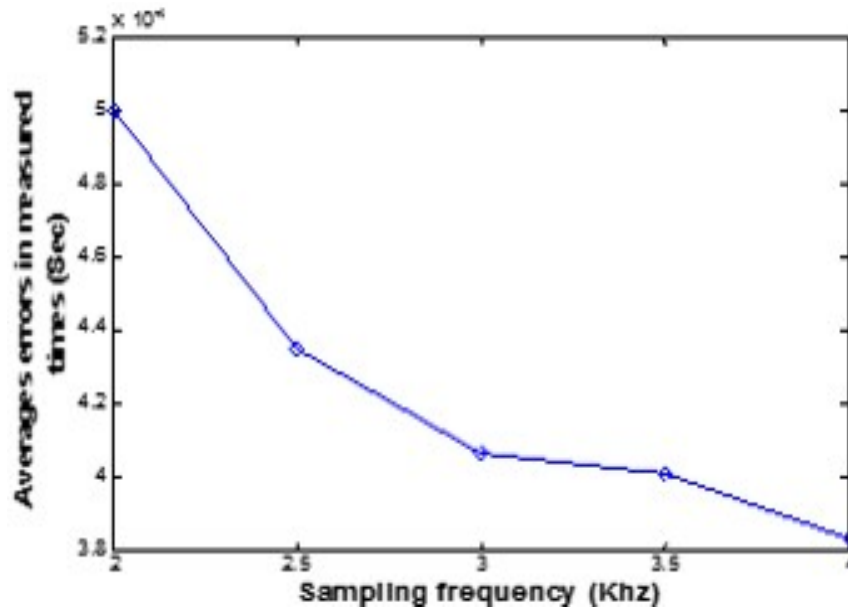


FIGURE 5.4: The impact of the sampling frequency on the uncertainty in TDOA measurements

a Gaussian model of a non-zero mean for relatively higher sampling rate, or a uniformly distribution model for small sampling frequency and short intervals between re-synchronisation events. Figure 5.5 and Figure 5.6 presents the variation in the form of the errors distributions in two principal cases respectively: a normal curve fit in the case of high sampling frequency and the form of uniform curve fit in the case of low sampling rate.

5.4.3 Estimation of the drift rate between notes

Equation 5.13 shows that for higher values of the term T , the right term can be written as:

$$\sqrt{2((1/f^2) + (T * \delta drift + \delta ub)^2)} \approx \sqrt{2 * (T * \delta drift)^2} \quad (5.14)$$

Thus, equation (5.13) becomes:

$$\delta_{t_{ij}} = \sqrt{2} \delta drift T \quad (5.15)$$

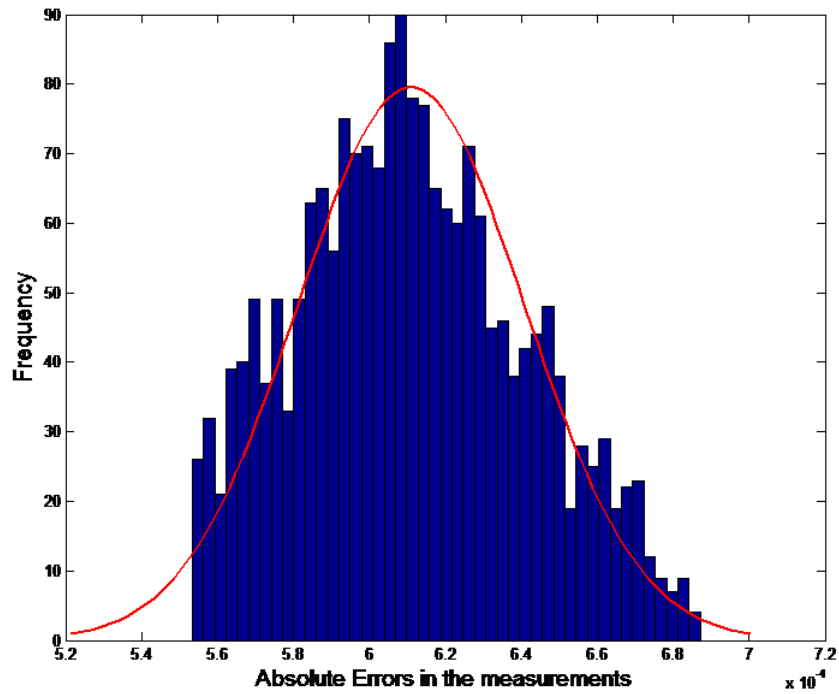


FIGURE 5.5: fig:Absolute errors distribution at re-synchronisation time of $T = 30$ sec, $\delta drift = 1.3 \times 10^{-6}$, δub taken to be equal to $1.3 \times 10^{-6} sec$ and the sampling frequency $f = 4.5 kHz$

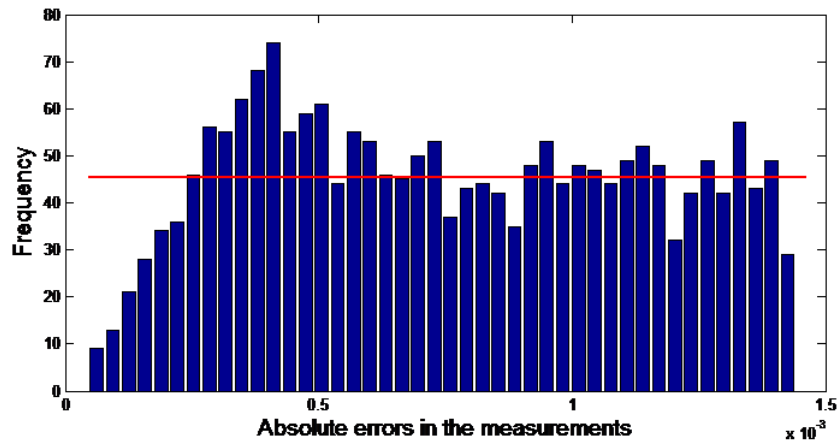


FIGURE 5.6: fig:Absolute errors distribution at resynchronisation time of $T = 20$ sec, $\delta drift = 1.3 \times 10^{-6}$, ub taken to be equal to $1.3 \times 10^{-6} sec$ and the sampling frequency $f = 1 kHz$

The estimation of the drift rate between the two sensors nodes is possible from the obtained linear equation in (5.15). This can be achieved by a set of experiments in which an acoustic event is emitted from a sound source to two equally distant sensors nodes from the source. These acoustic events have to be emitted at

gradually increased intervals of times. While the latter (the intervals of time) should be recorded with the corresponding time drift between these two sensors and therefore a graph that represents this linear relationship can therefore be drawn. Figure 5.7 presents the linear development of time difference between two sensors notes over time. It clearly shows how errors in measurement can take large magnitude without re-synchronisation.

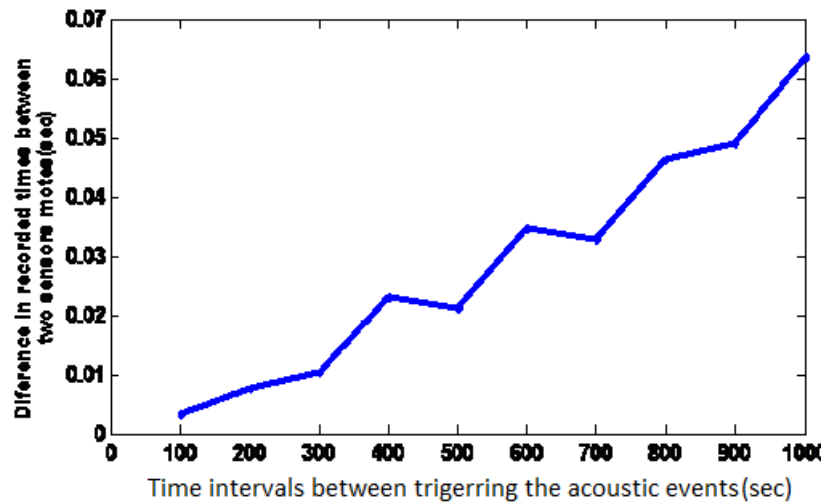


FIGURE 5.7: Development of the drift between two acoustic sensor nodes

Estimation of the drift rate between two sensor nodes is obtained from the slope of the linear Equation (5.15).

By using this experimental tool to estimate the drift rate ($\delta drift_{ij}$) between each pair of sensors i and j , by considering that the set of the sensors used has identical known values of acoustic sampling frequency, and by neglecting the effect of (ub), more accurate estimation of the matrix of uncertainties Σ (given in equation (5.7)) is obtained.

5.5 Powells Dogleg/Double Dogleg optimisers

As the problem in (5.5) does not have a close-form solution, a numerical solution has to be recursively obtained. Using the Gauss-Newton method to find a maximum likelihood estimate of the solution \hat{s}_{ML} is given by:

$$\hat{s}_{ML} = \underset{s}{\text{ArgMin}}(f^T(s)Wf(s)) \quad (5.16)$$

with $W = \frac{1}{\Sigma}$

A major disadvantage of the Gauss-Newton algorithm is its vulnerability to divergence unless it is initialised sufficiently close to the solution. To make up for this issue, a trust region based method called Powell's Dogleg is investigated in this work. This optimisation method was previously investigated in computer vision related problems [131]. The method works by combining both the Gauss Newton and the Steepest Descent directions steps. These are controlled explicitly via a radius Δ , called the trust region radius. The principle of this method is to find an approximation to step $S_{dl} = [\Delta x, \Delta y, \Delta z]^T$, (called the Dog Leg step), that guarantees the product $S_{dl}^T * S_{dl}$ to be inferior to Δ^2 . For f , given in (5.5), the Gauss-Newton step S_{gn} is estimated at every iteration using the least squares solution to the linearised system given by:

$$J^T J S_{gn} = J^T f \quad (5.17)$$

With J represents the Jacobian of f and given by:

$$J = \begin{bmatrix} \frac{\delta f_1}{\delta x}(s) & \frac{\delta f_1}{\delta y}(s) & \frac{\delta f_1}{\delta z}(s) \\ \vdots & \vdots & \vdots \\ \frac{\delta f_n}{\delta x}(s) & \frac{\delta f_n}{\delta y}(s) & \frac{\delta f_n}{\delta z}(s) \end{bmatrix} \quad (5.18)$$

The steepest descent step is given by a direction and a step size. The direction is defined as the function of the gradient that is given in the following form:

$$S_{sd} = -g \quad (5.19)$$

With:

$$g = J(s)^T f(s) \quad (5.20)$$

While the step size given by α is equal to:

$$\alpha = -\frac{(S_{sd}^T J(s)^T f(s))}{\|J(s)S_{sd}\|^2} = \frac{\|g\|^2}{\|J(s)g\|^2} \quad (5.21)$$

The selection of this value is justified by the need to reach a minimal value for the function f written in the form of:

$$f(s + \alpha S_{sd}) = f(s) + \alpha J(s)S_{sd} \quad (5.22)$$

The Powell's Dog Leg method combine the two candidate steps αS_{sd} and S_{gn} at every iteration to perform an estimation of the step size S_{dl} using the strategy

described in the following algorithm:

```

if  $\|S_{gn}\| \leq \Delta$  then
  |  $S_{dl} = S_{gn}$ ;
else
  | if  $\|\alpha S_{sd}\| \geq \Delta$  then
  | |  $S_{dl} = \frac{\Delta}{\|S_{sd}\|} S_{sd}$ ;
  | else
  | |  $S_{dl} = S_{sd} + \beta(S_{gn} - S_{sd})$ ;
  | end
end

```

Algorithm 1: Step size estimation in the DL Algorithm

An appropriate value of $\beta \in [0, 1]$ is chosen to insure $S_{dl} \leq \Delta$. However, by putting $a = \alpha S_{sd}$, $b = S_{gn}$ and $c = a^T(a - b)$ and after introducing a function ψ with:

$$\psi(\beta) = \|a + \beta(b - a)\|^2 - \Delta^2 \quad (5.23)$$

An optimal value for β can be obtained which ensure $\psi(\beta) = 0$. The development of ψ is given by the following:

$$\psi(\beta) = \|b - a\|^2 \beta^2 + 2c\beta + \|a\|^2 - \Delta^2 \quad (5.24)$$

From this equation, it is clear that ψ is a form of a second degree polynomial which is infinitely positive for β infinitely negative. Additionally, for $\beta = 0$, we have:

$$\psi(0) = \|\alpha S_{sd}\|^2 - \Delta^2 < 0 \quad (5.25)$$

which leads to the conclusion that ψ has one negative root. Additionally for $\beta = 1$, we have:

$$\psi(1) = \|S_{gn}\|^2 - \Delta^2 > 0 \quad (5.26)$$

Thus, ψ has a second root in $]0, 1[$. A most accurate computation of the positive root is given by the following pseudo code which ensure that $\|S_{dl}\| = \Delta$:

```

if  $c \leq \Delta$  then
     $\beta = \frac{(-c + \sqrt{(c^2 + \|b - a\|^2(\Delta^2 - \|a\|^2)})}{\|b - a\|^2};$ 
else
     $\beta = \frac{(\Delta^2 - \|a\|^2)}{(c + \sqrt{c^2 + \|b - a\|^2(\Delta^2 - \|a\|^2)})};$ 
end

```

Algorithm 2: Parameter estimation of the DL region

To control the size of the radius Δ of the trust region, this method adopts the gain ratio ρ . This is calculated as the following:

$$\rho = \frac{f(x) - f(x + S_{lm})}{L(0) - L(S_{lm})} \quad (5.27)$$

With L is a linear model introduced to insure iteration monitoring. It is defined by:

$$L(S) = \frac{1}{2} \|f(S) - J(S)\Delta S\|^2 \quad (5.28)$$

A large value of ρ indicates that the linear model is good. Hence, the radius Δ is increased leading to taking longer steps that will be closer to the Gauss-Newton direction. However, if ρ is small (or even negative), then the value of Δ will be reduced, implying taking smaller steps that are closer to the steepest descent direction. Considering ub , lb , and ϵ as user defined parameters, the update of the radius Δ is completed according to the following strategy (Algorithm 3):

```
if  $\rho > ub$  then
   $\Delta := \max(\Delta, 3 * \|S_{dl}\|);$ 
  if  $\rho < lb$  then
     $\Delta := \frac{\Delta}{2};$ 
     $found := (\Delta \leq \epsilon(\|s\| + \epsilon));$ 
  end
end
end
```

Algorithm 3: Trust region Radius Estimation

In the following is the overall steps of the DogLeg algorithm applied on a least square scheme.

```

 $k := 0; s := s_0; \Delta := \Delta_0, W = W_0;$ 
 $ub = 0.8; lb = 0.25; g := J(s)^T W f(s);$ 
 $found := ((\|f(s)\|_{\text{inf}} \leq \epsilon_3) \text{or} (\|g\|_{\text{inf}}));$ 
while (not found) and ( $k \leq k_{\text{max}}$ ) do
  |  $K := k + 1;$ 
  |  $\alpha = \frac{\|g\|^2}{\|J(s)Wg\|^2};$ 
  |  $h_{sd} = -\alpha g;$  compute  $S_{gn}$  ; call Algorithm 01 to compute  $S_{dl};$ 
  | if ( $\|S_{dl}\| \sec \epsilon_2 (\|s\| + \epsilon_1)$ ) then
  | | found :=true;
  | else
  | |  $S_{\text{new}} = s + S_{dl};$ 
  | | Compute  $\rho$  according to (14);
  | | if ( $\rho > 0$ ) then
  | | |  $s = s_{\text{new}};$ 
  | | |  $g := J(s)^T W f(s);$ 
  | | |  $found := ((\|f(s)\|_{\text{inf}} \leq \epsilon_3) \text{or} (\|g\|_{\text{inf}}));$ 
  | | | call Algorithm 2 to update  $\Delta;$ 
  | | end
  | end
end

```

Algorithm 4: The Double Dog Leg Algorithm

5.5.1 Powell's method with Double Dogleg step

Aiming to improve both the speed and the overall accuracy of the Powell's Dog Leg method, we adopted its Double Dog Leg version. In the original method, the optimal trajectory follows the steepest descent direction to the Cauchy point

before it converges to the Newton point forming a Dog Leg step (Figure 5.8). This step should be intersecting with the trust region boundary defined by the radius Δ . By introducing an intermediate Gauss Newton step between the Cauchy Point and the actual Newton point, a change in the behaviour of the Powell's Dogleg algorithm is expected. This change offers a further improvement so that the new optimal curve trajectory crosses the trust region boundary earlier than the original method giving a faster optimisation. Hence, this method is called the double Dogleg algorithm,[128, 132]. The new trajectories are ruled out by adjusting the following equation given in Algorithm (1).

$$S_{dl} = S_{sd} + \beta(S_{gn} - S_{sd}) \quad (5.29)$$

To be written as:

$$S_{dl} = S_{sd} + \beta(S_{gn} - \lambda S_{sd}) \quad (5.30)$$

With $\lambda = 0.8$ is the optimal value that ensures higher convergence rate.

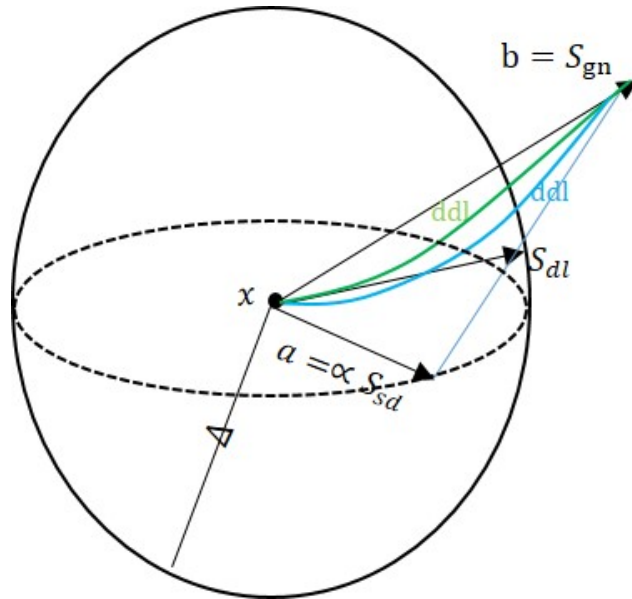


FIGURE 5.8: The Powells Dogleg and Double DogLeg step

5.5.2 Weighting norms adoption

The original Powell's Double Dogleg method adopts the linear least squares optimisation framework and the standard steepest descent steps to estimate the Double Dogleg step. Considering the uncertainties in the measurements and taking them into account in the optimisation process is aimed to increase the overall accuracy of this method. This can eventually be achieved using the weight least squares framework and the weighted steepest descent in Equation (5.17) and (5.19) respectively.

The acoustic measurements that are more informative are given lower weights compared to the less informative ones. By adopting a weighed least square form (WLS) for estimating the Gauss-Newton Step in Equation (5.17), the new step will be estimated from the following:

$$J^T * W * J * S_{gn} = -J^T . W . f \quad (5.31)$$

In analogy to (5.31), the weighted steepest descent step is given by the following:

$$\alpha S_{sd} = \alpha J(s)^T . W . f(s) \quad (5.32)$$

With W are the weights obtained from the uncertainties in the measurements as explained in Equation 5.4.

With the changes made to the Gauss-Newton step and the steepest descent step including the uncertainty norms, the Double Dogleg algorithm is made robust.

5.5.3 The total least squares (TLS)

While the name total least squares has appeared only recently in the literature [123–125], this method is not new and has a very long history in statistical literature. It was known by different names such as orthogonal regression, errors-in-variables, and measurement errors. The univariate problem ($n = 1, d = 1$) has been discussed already in 1877 in [126]. While about thirty years ago, the technique was extended in [127] and in [128] to multivariate ($n > 1, d > 1$) problems. We introduce this method to the problem for estimating the Gauss-Newton step.

By having $A = J^T W J$ and $B = -J^T W f$, with J and f are the Jacobian and the function given in (5.5) and (5.18) respectively. The problem of Gauss-Newton step estimation can be written then as :

$$A.S = B \quad (5.33)$$

The least squares method reaches the solution after making correction to term B while term A remains unchanged. The total least squares suggests that since both B and A are input data they can be treated symmetrically. That is why it seeks minimise (in the Frobenius norm sense) corrections ΔA and ΔB of the given terms that make the corrected system in Equation (5.33) of equations solvable:

$$\begin{cases} \hat{A}S = \hat{B}, \\ \hat{A} = A + \Delta A, \\ \hat{B} = B + \Delta B, \end{cases} \quad (5.34)$$

The total least squares solution is estimated using the singular value decomposition of C with $C = [AB]$. Having $C = U\Sigma V^T$, where $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_{n+d})$ is a singular value decomposition of C , $\sigma_1 \geq \dots \geq \sigma_{n+d}$ is the singular values of C ,

and defining the partitioning:

$$V = \begin{matrix} & n & d \\ \begin{matrix} n \\ d \end{matrix} & \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \end{matrix} \quad (5.35)$$

and

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \quad (5.36)$$

A total least squares solution exists if and only if V_{22} is non-singular. In addition, it is unique if and only if $\sigma_n \neq \sigma_{n+1}$. In the case when the total least squares solution exists and is unique, it is given by:

$$\hat{S}_{tls} = -\frac{v_{12}}{v_{22}} \quad (5.37)$$

with these changes, the algorithm is brought into being a robust tool of estimation as described in section [5.5.1](#), [5.5.2](#) and [5.5.3](#).

5.6 Experimental setup and experiments

To evaluate the performance of the presented acoustic source localisation approach in comparison with previously investigated methods in this area; we used a WSN composed of a set of 04 MICAZ motes, Figure [5.9](#), with the MTS310 sensors boards, Figure [5.10](#). This was the minimum necessary setup to ensure a 3D localisation. The sensors nodes were deployed in a space of dimensions 2.5 m (W) x 4.5 m (L) x 2.5 m (H), Figure [5.11](#). Each sensor was positioned at each corner of the area with the up side pointing to the centre of the space.

An acoustic pulse test of a frequency 4 kHz is used. It is played through a sound buzzer at different points in the designed space of the experiments. The



FIGURE 5.9: Micaz sensor mote



FIGURE 5.10: The Mts310 sensor board

pulse was experimentally selected in order to generate a reasonable pulse shape for the conducted experiments. The sensors were periodically receiving reference broadcast synchronisation (RBS) messages from a fifth sensor before realising the sound. As soon as the burst of sound is detected by the microphone using a thresholding mechanism, the related detecting times are recorded before being sent to a base station in a TinyOs message. The base station is another MICAZ mote plugged to the gateway board MIB520 where the received measurements in the hexadecimal form are processed with the proposed acoustic source localisation algorithms. For an accurate evaluation of the proposed algorithm, a high number of test points is required.

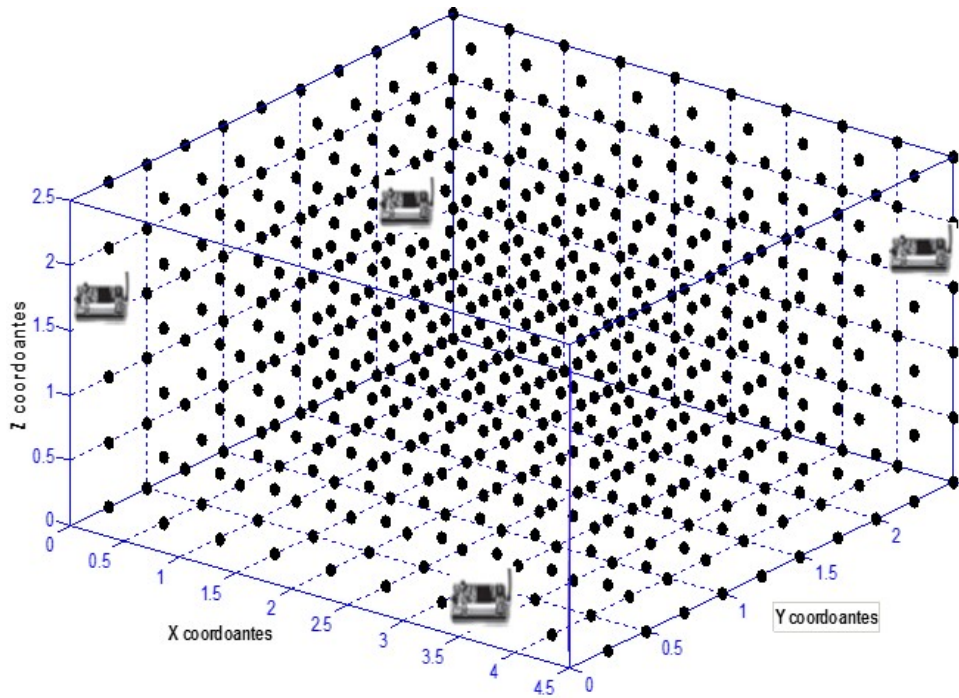


FIGURE 5.11: Sensors and source location used to evaluate the performance of the proposed method

5.6.1 Performance evaluation of the the proposed method

By comparing the Euclidean distance between the ground truth and the estimated sound location with a given threshold that represents the tolerated errors in localisation, we can evaluate a given algorithm and decide whether it can reach a targeted level of accuracy. Comparative results of the overall accuracy between the Double Dogleg and Gauss-Newton is depicted in Figure 5.12 (with threshold of $th = 50cm$).

It presents clearly the overall merit of the techniques based on the Powell's methods (Dogleg/Double Dogleg based) in converging to the global minimum for the acoustic source localisation problem. Comparing this convergence rate with the one achieved by the classical Gauss-Newton method, demonstrates very well our motivation behind adopting such an optimisation technique. It worth noting that

this high convergence success rate is due to the Dogleg/Double Dogleg optimisation technique and not to the scheme of optimisation (least squares, total least squares,...etc). The latter is based on least squares scheme for both dogleg/double dogleg and Gauss-Newton techniques.

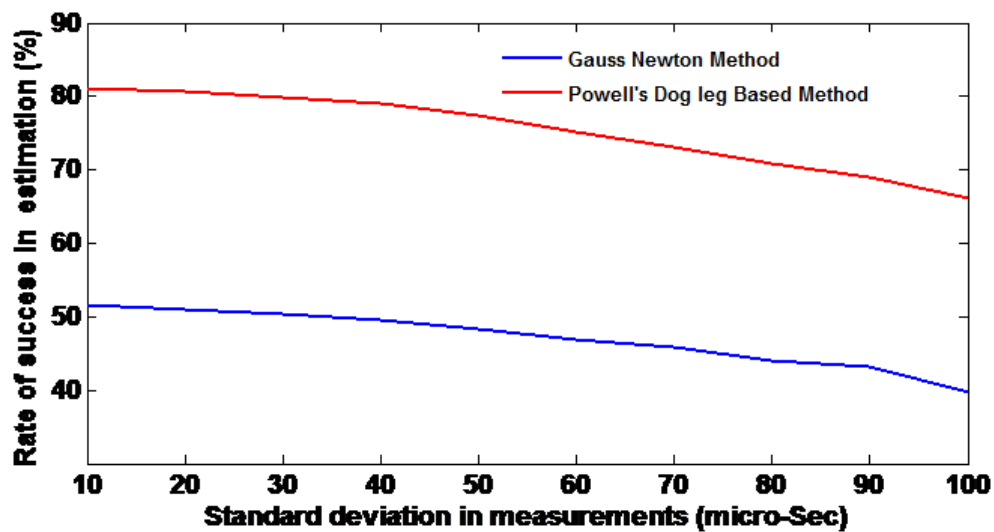


FIGURE 5.12: Evaluation of the overall accuracy of the family of the trust region method and the linear search by Gauss-Newton method

The performance of the different algorithms has been compared using the Root Mean Square Errors (RMSE) between the ground truth and the estimated position for 200 test experiments. This measure was calculated at increased level of errors in the measurements that were controlled by varying the sampling rate of the sensor nodes. The different versions of the Powells method (Dogleg, Double Dogleg) show their superiority in comparison with the Gauss-Newton methods (with least squares (GNLS) and weighted least square (GNWLS) schemes). Indeed, Figure 5.13 shows how the localisation errors are minimised in the different versions of Powell's methods while varying the sampling frequency. As shown in Figure 5.4, by varying the sampling frequency, the level of uncertainty acting on the TDOA measurements varies as well.

Figure 5.13, nicely explains how it is important taking into account the uncertainty in a weighted Total Teast Squares scheme with the Double Dogleg optimisation technique to achieve best localisation results. Furthermore, Figure 5.14 shows that Double Dogleg Total Least Squares (DDLWTLS) algorithm provides the best error minimisation result when a sample frequency of 4.5KHZ is considered. This sample frequency induces a small amount of uncertainty. However, varying the time intervals between re-synchronisation events will incur additional uncertainties as presented in Figure 5.3. Double Dogleg Weighted Least Squares (DDLWLS) introduced more robustness than Double Dogleg using only Least Squares scheme. In Figure 5.15, the sampling frequency considered is 1.5KHZ. This frequency in addition to varying the time intervals between re-synchronisation events will induce more uncertainty than the case considered in the experiment of Figure 5.14. This augmented uncertainty increases the magnitude of errors for all the studied localisation algorithms as shown in Figure 5.15. However, DDLWTLS is showing coping very well comparing with all other algorithms.

5.6.1.1 Convergence rate and execution time

A numerical evaluation of the convergence rates of the presented methods is achieved by running the implemented algorithms for some observations in a situation of a zero residual. Considering the norms of the residuals on each iteration plots those convergence rates. The latter are presented by the function CR as following:

$$CR(k) = \|S_k - S^*\|_2 \quad (5.38)$$

With S^* is the optimal localisation solution, while S_k is the estimated position at iteration k . The plot of CR in Figure 5.16 shows the variation in the convergence speed of the different methods. The convergence rate for each method varies from iteration to another. Although the Gauss-Newton based methods converges

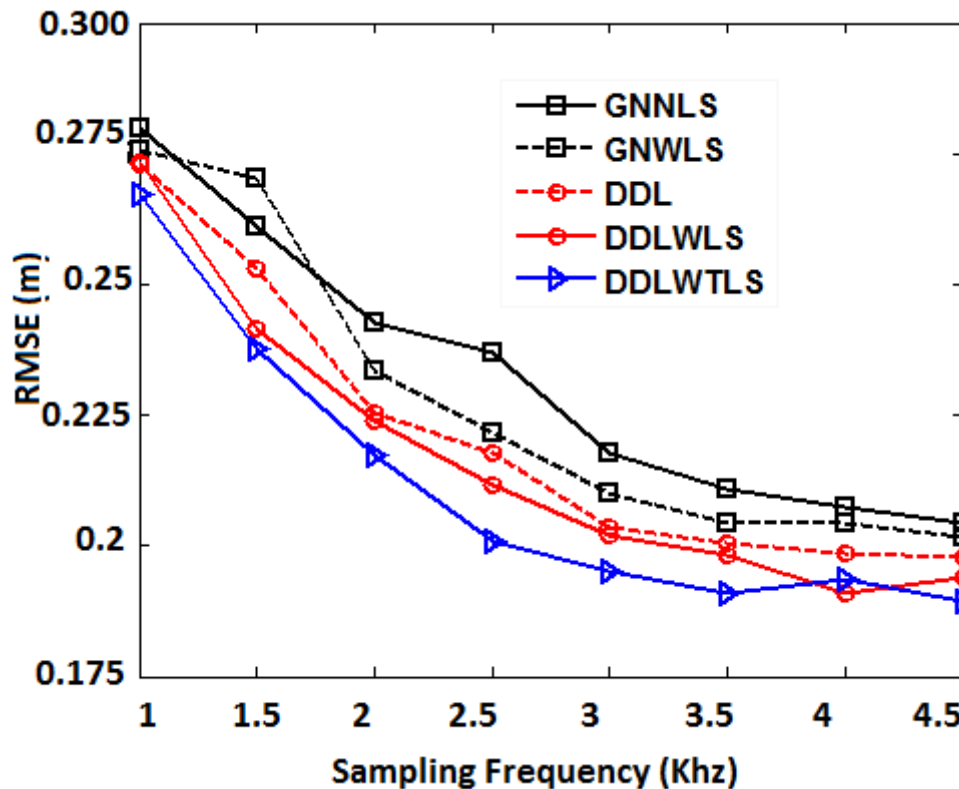


FIGURE 5.13: Evaluation of the overall accuracy of the double dogleg method at varying frequencies

initially into a better solution than Dogleg and Double Dogleg initial solutions, the later optimisation technique presents similar convergence behaviour than Gauss-Newton technique with even better slope of convergence taking into account their respective initial iteration solutions.

Having said that, the last finding resulted in a longer execution time for DDL-WLTS compared with to GNLS and GNWLS as shown in Figure 5.17. This execution time increases proportionally as the number of the deployed number of sensors is increased though this is a common trend in all the existing methods. It is also important to note that the overall execution time using the same number of sensors is shorter in the case of using the Double Dogleg technique with Least Squares scheme than Double Dogleg technique with Weighted Total Least Squares scheme. By increasing the number of the acoustic sensors in the arena

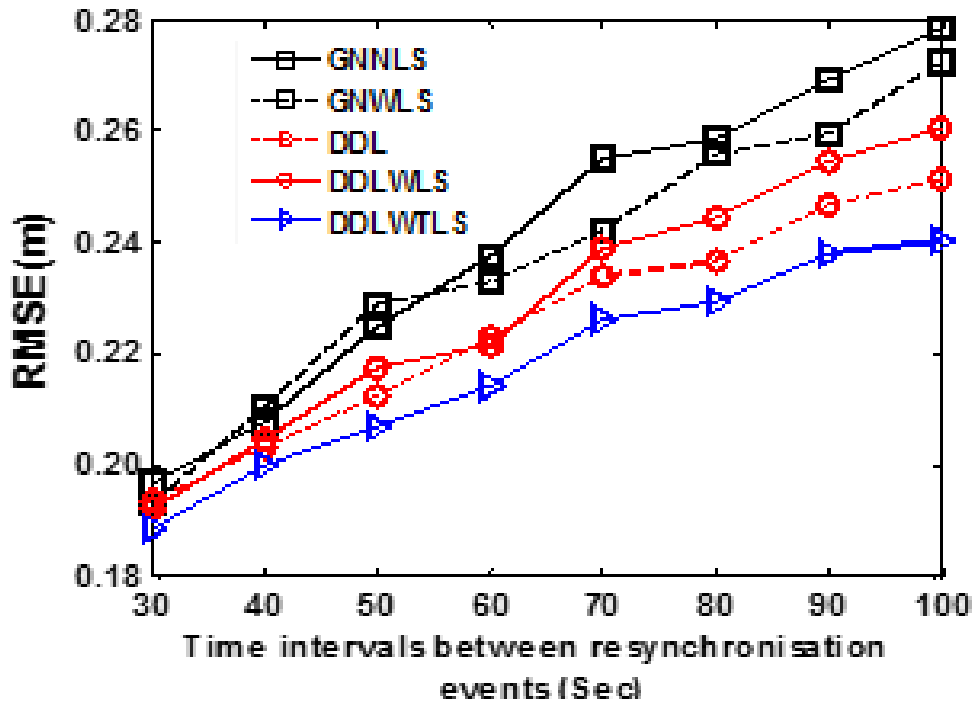


FIGURE 5.14: Evaluation of the overall accuracy of the double dogleg method at high frequencies

we achieve improved localisation accuracy for all the proposed methods in this study (Figure 5.18). A better accuracy is recorded for the DDLWTLS method though. We also noticed that going beyond the deployment of 16 sensor motes is not impacting on the acoustic source localisation accuracy anymore.

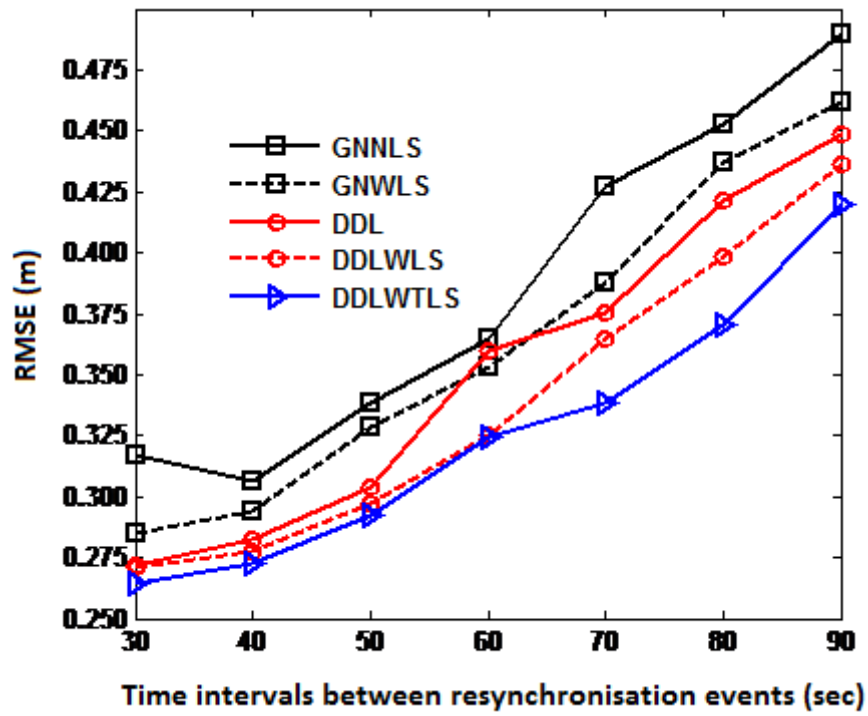


FIGURE 5.15: Evaluation of the overall accuracy of the Double Dogleg method at low frequencies

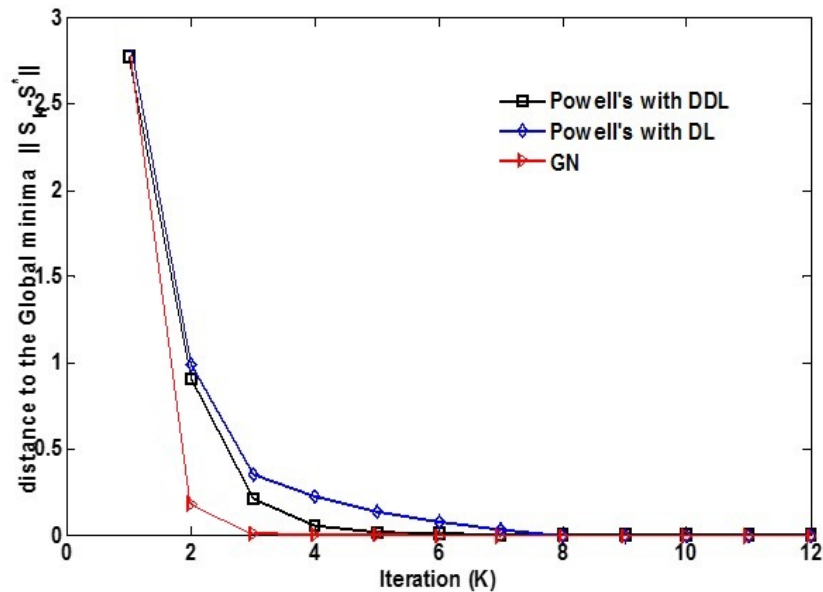


FIGURE 5.16: Convergence rate of the Powells double dog leg method over the gauss newton method

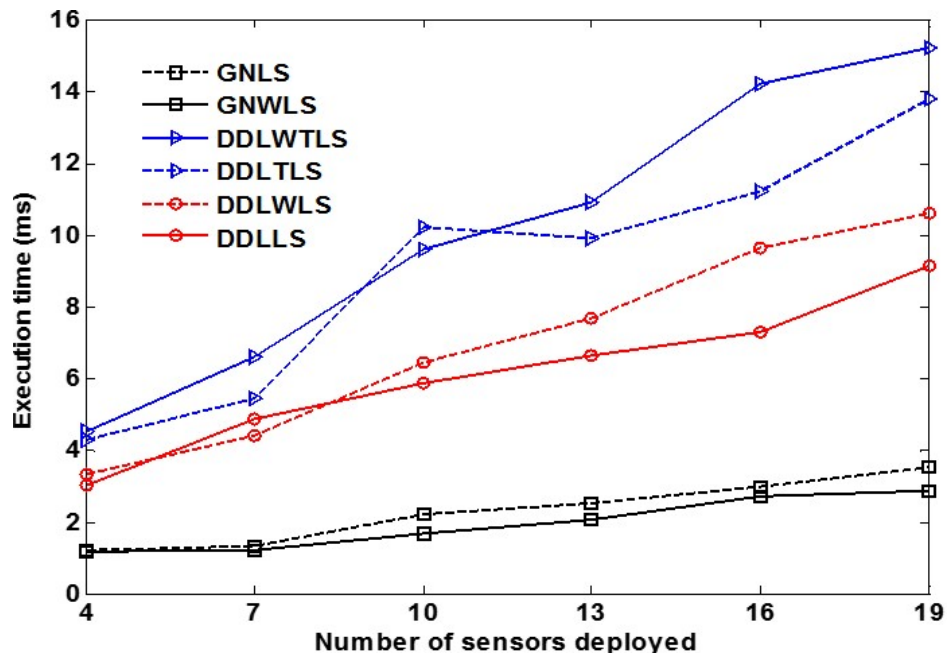


FIGURE 5.17: Execution time for the different methods

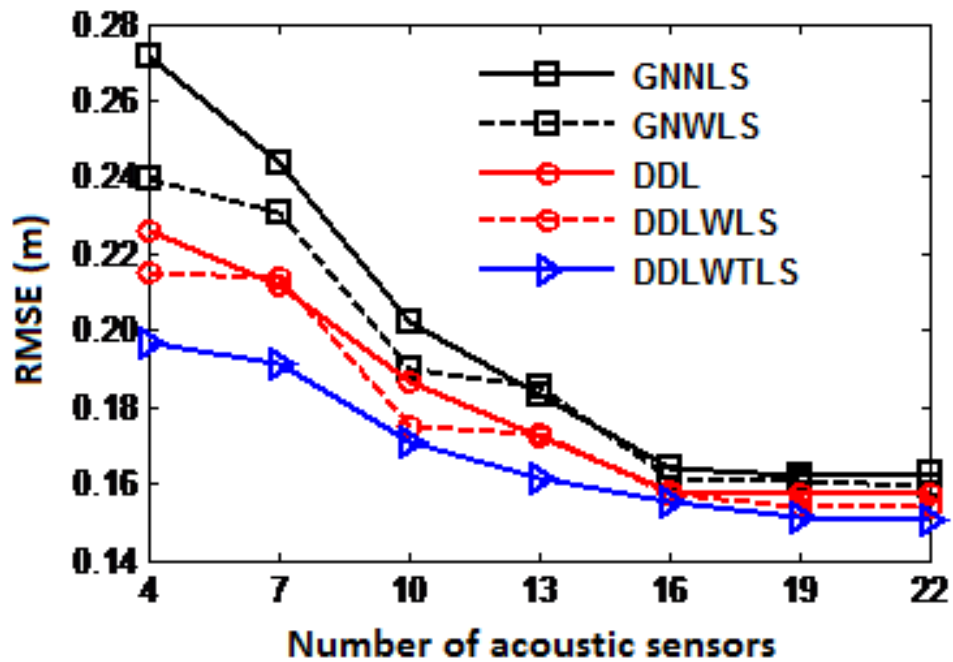


FIGURE 5.18: impact of the WSN size on localisation accuracy

5.7 Conclusion

We proposed in this chapter innovative Powell,s Dogleg and Double Dogleg optimisation techniques in dealing with the problem of acoustic source location in WSN. We showed the efficiency and accuracy of those techniques in comparison with a classical linear search based technique such the Gauss-Newton. The Double Dogleg technique combines the advantages of the Gauss-Newton technique, which offers rapid convergence near the solution and the steepest-descent method, which is robust and numerically stable far from the solution. Hence, with the adoption of this technique, a higher probability of convergence can be obtained by adjusting the trust region radius.

A proper experimental based formulation of the origin of uncertainties and their magnitudes biasing the accuracy of the TDOA measurements has been newly introduced in this Chapter. Dealing efficiently with those uncertainties in the acoustic source localisation based optimisation framework requires not only an efficient optimiser but also adopting the optimisation scheme, which provides the tools to do so. Indeed, accuracy improvements of the Powell's Double Dogleg method is obtained using a weighted norms representing uncertainties in a total least squares scheme to estimate the steepest decent and the Gauss-Newton steps.

This strategy might end by increasing a little the computational time of the estimation to a still reasonable level, but it is guaranteeing the convergence to a global minimum solution while dealing efficiently with the system uncertainty.

This page is intendedly left blank

Chapter 6

Active Acoustic Sources Localisation in Distributed Sensor Networks

6.1 Introduction

In this chapter we investigate the problem of improved detection and localisation in distributed sensor networks. In the field of motion detection and-or localisation, it is a common situation that movements of objects is accompanied by special emitted sound that varies in energy pressure, loudness or related features. Therefore, a basic approach is to imitate the cooperative functioning of human senses combining both vision and hearing capabilities for better detection and position estimation of moving objects. Within this context, we firstly propose an innovative solution for active acoustic sources detection and localisation in a distributed sensor network. This solution is based on augmenting the RGB vector, in the SGGMM background subtraction method proposed in Chapter 4, with the

acoustic information to detect possible moving sound sources. Secondly, we investigate the design of a centralised/decentralised architecture of fusion acoustic and visual data. The aim of this architecture is to improve the quality of tracking of active sound sources in a distributed sensor network.

6.2 Related works

Research in the domain of data fusion has gained a lot of interest in recent decades due to the diversity of its application and the ability of sensing technology to deliver outstanding results. This is regardless of its functioning whether it is in a cooperative, competitive or complementary manner. The problem of acoustic and video data fusion has been widely investigated, especially to deal with problems of speaker detection and localisation. In this context, references [133–135] have demonstrated the correlation between audio and video modalities in the speech case. They show that the correspondence between the speaker lips movements and the produced sounds can be exploited by the receiver to understand better the speech, especially in the presence of noise. Another example that demonstrated the relationship between hearing and vision in speech perception is the McGurk effect, described in [136]. This effect is generally experienced by a combination of a video of a person uttering one phoneme with a soundtrack corresponding to a different phoneme. Additionally, a new form of combination of the two data modalities is investigated in [137]. This form is based on the assumption that both the acoustic and the visual models can be estimated as part of a joint unsupervised optimisation for target localisation. In [138], a system was designed for detection and localisation of active speakers by combining the visual reconstruction using a stereoscopic camera pair and sound-source location using several microphones. For surveillance systems and tracking, this type of fusion was of high importance as shown in [139, 140]. In [139], a sensor fusion framework based on particle

filters was proposed. It aimed at combining both results of the detection and the tracking from a co-located acoustic array and video camera for vehicles tracking. The particle filter based trackers were used to recursively estimate the state probability density functions for the combined tracker. The overall performance of the target tracking was shown to be improved as the video controls the particles diversity at low signal-to-noise (SNR) levels of the acoustics. Results obtained in this work as well as in [140], were promising, showing the ability of the particle filter of tracking the change in target motion model without prior modelling. However, higher performance of this type of filter can be achieved only with a high number of particles, which results in overloading the system. Additionally, in [141] a solution based on a centralised/decentralised architecture using the extended Kalman filter (EKF) for the target dynamic model was proposed. It adopted a foreground-background modelling technique with a skin colour tracker for visual appearance tracking. For the acoustics, a microphone array with a beam former was used to locate the acoustic source.

In a similar context, a system was proposed in [142] which aimed for ships identification and localisation based on the fusion of acoustic and the video data. In this approach, the fusion enabled the estimation of sound attenuation in a wide frequency band and the collection of a noise library of various ships. The latter was used for ship classification by passive acoustic methods. Similarly, in [143] a description of a knowledge-based system designed to detect evidence of aggression by means of audio analysis and camera sensors network was presented. When the aggression event was detected, the images captured by the networked cameras were sent to a central system for further analysis and decision making.

In this chapter, we contribute in its first part to the problem of improved detection of active acoustic sources using a cost effective fusion scheme that works in distributed sensor networks. The proposed solution suggests including the information corresponding to the location of the acoustic source in the image frame.

This is processed using a background/foreground segmentation method such that only moving objects with sound activity are included in the foreground.

In the second part of this chapter, we evaluate the performance of active acoustic sources localisation and tracking using a centralised/decentralised architecture, in which we compare the performance of two classical fusion algorithms. The first is based on the Covariance Intersection (CI) [144, 145] while the second is based on the Information Fusion (IF). We also compare the accuracy of tracking using the centralised/decentralised architecture to the accuracy obtained using one single type of data (acoustic or visual).

6.3 First part: Augmented SGGMM with the acoustic information

In this section, the problem of detection and localisation of active sound source is investigated using a new fusion approach. The proposed solution aims to combine two data modalities by augmenting the 3-D vector of RGB colours utilised by the Spatially Global Gaussians Mixture Model (SGGMM) proposed in chapter 4 with the acoustic information. By using this fusion method, we look at improving the detection accuracy of moving acoustic sources. Indeed, evaluation results using an implementation of this fusion scheme on a distributed sensor network showed detection improvement compared to using the SGGMM based on vision only. Moreover, the technique permitted reaching higher localisation accuracy of moving sound sources in comparison to using acoustic measurements only.

6.3.1 Proposed Fusion architecture

For this fusion approach, we deal with heterogeneous distributed sensors network composed of a smart camera (the CITRICc) and seven (07) Micaz sensor nodes.

The Micaz motes (described in Chapter 3) can communicate with the camera board through the Telosb mote. The latter is used for energy supply while playing the role of a mediator with the external world for the camera using the active messaging [146]. Note that a second communication technique using the serial packets, is used for ensuring communication between the camera board and the Telosb. The sensors network were set up in an indoor environment (Heaviside Laboratory at Cranfield University) where the tests have been conducted. The testing was limited to image sequences of resolution (320×240) to ensure the shortest processing time.

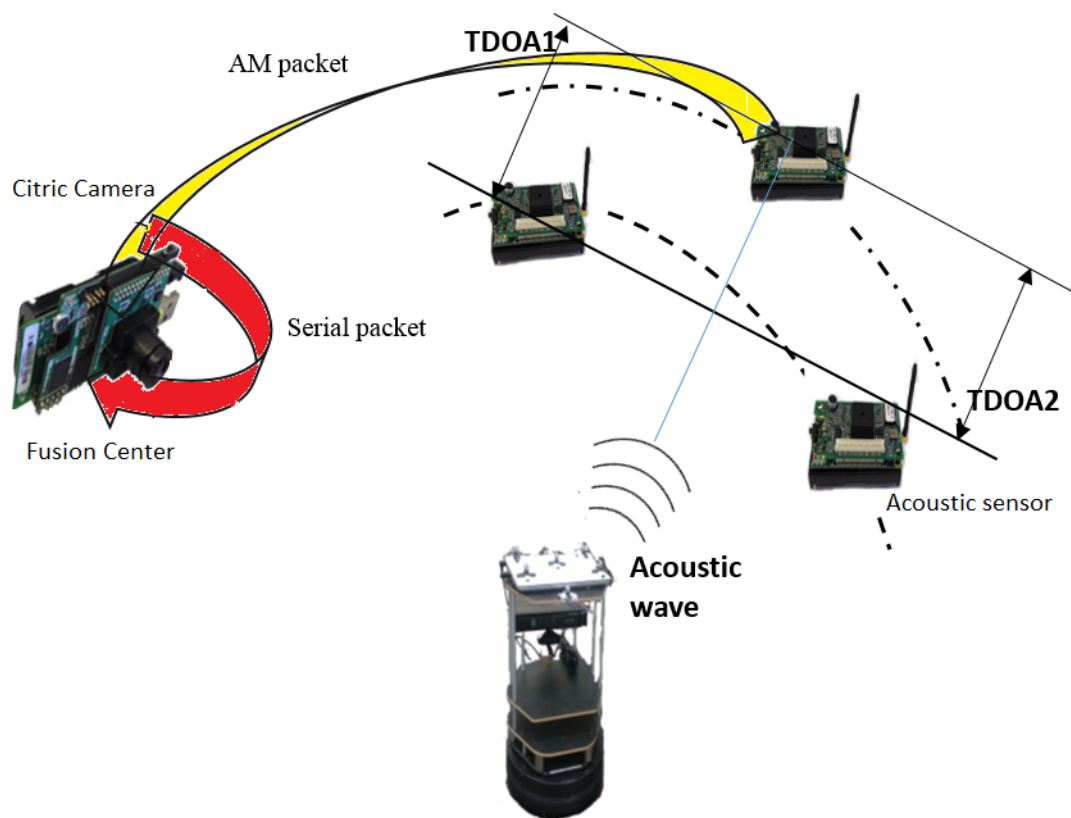


FIGURE 6.1: Different components of the fusion architecture proposed

For software implementation at the CITRIC camera node, the SGGMM method (described in Chapter 4) was used. When the camera receives time measurements of the acoustic events in the scene, this information is included in the augmented

SGGMM model. The acoustic information follows different stages before it becomes ready to be utilised in the image information:

- Firstly, synchronisation between acoustic sensors should be ensured. This task is completed using the broadcast reference synchronisation technique [147], in which reference beacons are sent periodically from a coordinator (Synchroniser) mote (a Micaz mote). This is done because the programming tool (NesC under TinyOs), used to build the codes for the sensor motes, is deprived of any local time synchronisation capability.
- When the acoustic sensors (situated at different locations) receive the synchronisation message, their local timers are triggered and they start listening for any sound of loudness that exceeds a pre-defined threshold. In case of positive detection, the corresponding times of detection are sent to the CITRIC camera for further processing.
- The received times at the camera node are used to compute the TDOA measurements. These are employed later to estimate the position of the detected moving object using the Double Dogleg method, developed in chapter 5. Through projection of the statistics of the sound position to a duplicate of the image frame, using a projection function that uses the camera calibration parameters, we obtain finally what we call an acoustic channel. The latter is concatenated with the RGB images captured by the camera. From these built 4D images, a foreground removal step is computed to extract pixels belonging to the active acoustic sources.

Figure 6.2 gives a general description of the main modules used in this programme as implemented in the CITRIC camera. Details of the implemented application for the acoustic localisation using the sensors mote is given in section 6.3.2. Section 6.3.3 presents solutions to some issues related to the acoustic localisation in addition to details of augmenting the SGGMM model to include the acoustic source location.

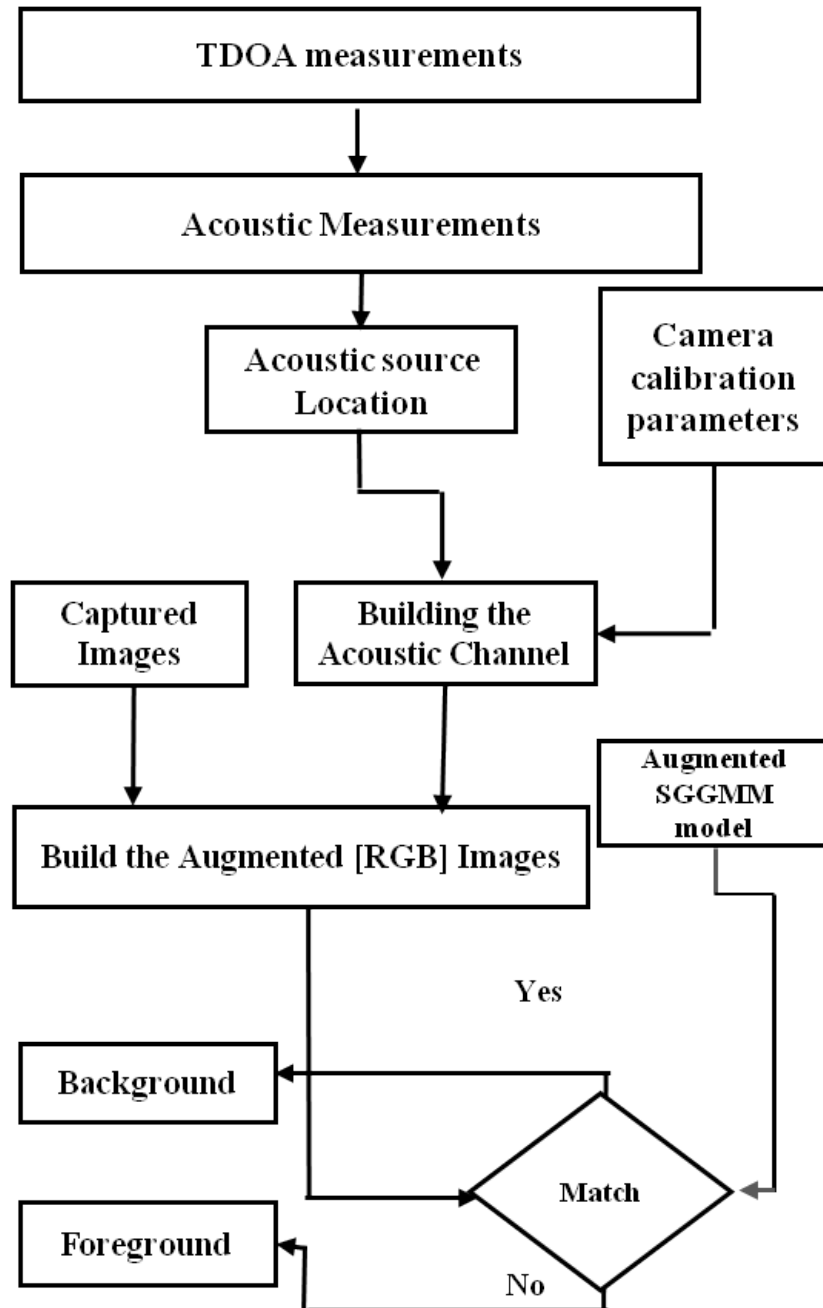


FIGURE 6.2: Software architecture of the proposed solution implemented at the CITRIC camera

6.3.2 Time of acoustic events measurements using the Micaz sensor motes

Due to the limited signal processing capabilities of Micaz motes, we focussed in this study on measuring some special values related to the variation of the signal energy. In the proposed method, similar to the one presented in [146], the signal energy is compared to a given pre-defined threshold. In case of an event of a loud sound, the time when the energy values exceed this threshold is recorded and send to the CITRIC camera. To this end, a programme [148] that ran in Micaz mote for loud sound event detection was designed using NesC under the TinyOS. Programming in NesC involves the creation and the wiring between different modules and interfaces [146]. Major components of the programme implemented at the acoustic sensor level are shown in Figure 6.3-(a). It is composed of a main module named **Coordinator**. The latter is connected to a module named **Detector** which is responsible of triggering an alarm when the microphone output value exceeds a given threshold. The module called **Microphone** is used to ensure control over the microphone input and output (warming, setting up the gain, reading).

The Coordinator component is also connected to the following standard interfaces: the **Radio**, which handles communication with other motes; the **Timer** (the local clock) and the microphone interface that is responsible for controlling the power of the microphone. A second algorithm is implemented on the Synchroniser mote to ensure the synchronisation between the different acoustic sensors. The Synchroniser broadcasts a reference beacon to the neighbouring acoustic sensors. This algorithm is centred on a main module **Synchroniser** that is wired to two system interfaces(Radio and timer) as shown in Figure 6.3-(b). Provided that all the sensors are within radio range, when acoustic sensor motes receive the synchronisation message, their timers are initialised, following by warming up the microphone controller.

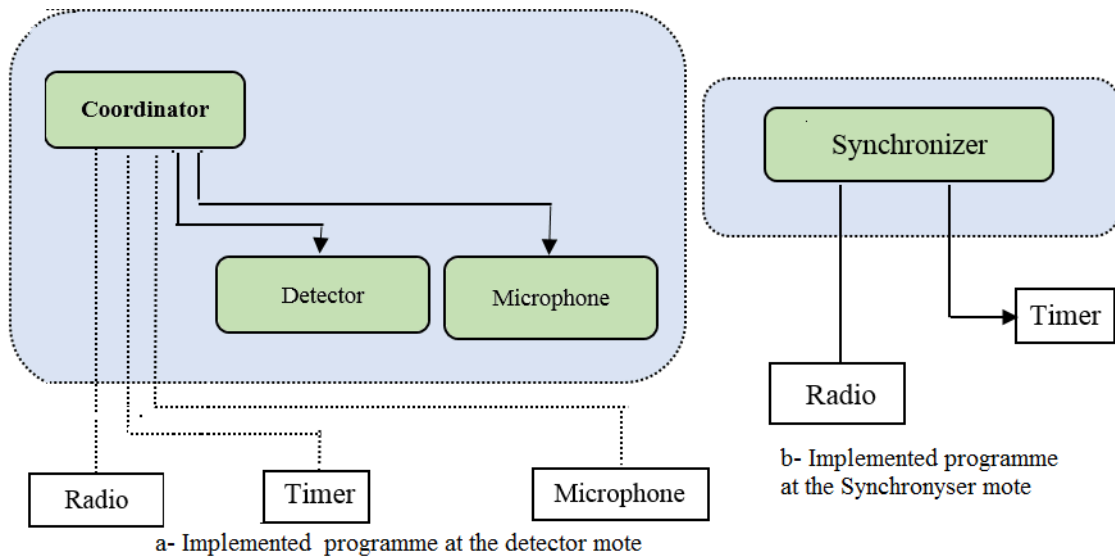


FIGURE 6.3: Architecture of the Nesc programme in charge of acoustic event detection

For the conduct of the experiments, the acoustic sensors had to be distributed in such a way to permit a reliable detection of the energy bursts. This is ensured by placing the sensors at distant position from the wall to ensure free field conditions. Additionally, the face of the microphones were installed facing the top to increase the sensitivity.

6.3.3 Acoustic source localisation with outliers and erroneous measurements elimination

In order to build an acoustic channel that can be combined with the RGB images using the SGGMM model, the received times at the CITRIC camera are firstly converted to TDOA measurements. We have shown in Chapter 5 the performance of the trust region method (the Double Dogleg weighted least square DDLWLS) in estimating the position of the acoustic source based on the optimisation of a non-linear least square problem. This method is adopted in this chapter.

In a TDOA based acoustic localisation, an accurate estimation of the acoustic source using a non-linear least square approach depends heavily on the accuracy of the related TDOAs measurements. However, the presence of outliers and errors in such measurements is inevitable. This is mainly due to the nature of the acoustic signal propagation and components reliability of the acoustic sensor nodes. Figure 6.4 shows how an erroneous measurement of one sensor in a sensor network composed of seven (7) nodes can lead to a large error in the localisation. Indeed, for such a number of nodes, an error of 8 ms in one sensor can lead to a solution of about 6 m far from the real position.

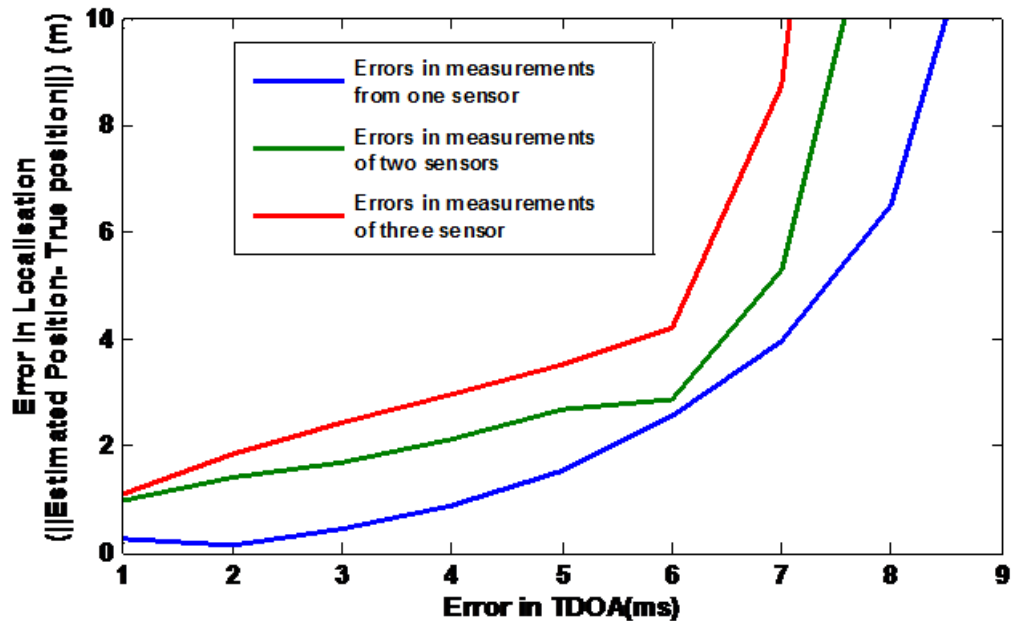


FIGURE 6.4: Propagation of localisation errors in a distributed acoustic sensors network composed of 7 nodes with erroneous measurements.

To deal efficiently with this problem, we adopted a solution based on the Random Sample Consensus (RANSAC) algorithm, firstly proposed by [149], and adopted for acoustic problems in [150]. For robust acoustic localisation in our distributed sensors network, the solution works through the following steps:

- 1 Initialisation: $k = 1$; n_O : maximum number of TDOA measurements.

- 2 Randomly select a set of three TDOA measurements, check for $i = 2, 3$ that $|TDOA_1 - TDOA_i| < \epsilon$ (with ϵ chosen to be relative to the maximal space between two sensor nodes), otherwise go to step (5).
- 3 Estimate the source position \hat{r}_s using the Double Dogleg method.
- 4 Among the remaining TDOAs measurements, compute the theoretical TDOA $\hat{\tau}_i, i = 4, \dots, 6$ with respect to the estimated position \hat{r}_s . A record of the persistent set S_k of $\hat{\tau}_i, i = 4, \dots, n_O$ should be made for each time difference of arrival satisfying $|TDOA_i - \hat{\tau}_i| < \epsilon$
- 5 Increment: $k = k + 1$, if $k \leq n_O$, then goto step (2).
- 6 Among the persistent sets $S_k, k = 1, \dots, n_O$, select S^* that has the maximal number of TDOA measurements. The estimated source position with the TDOAs in the set S^* is selected to be corresponding to the target of interest.

6.3.4 Dealing with time delay of measurements arrival

Due to hardware and physical issues (packet loss, sound events emitted at separate frame times), the acoustic information arrives at a slower rate compared to the images. Therefore, predicting the position of the acoustic source location when the related information is absent remains a necessary task. However, since the change in position of the active acoustic source is not noticeable during the camera frame processing ($\Delta t = 0.33$ sec), this position is assumed to be fixed. Hence, for ease of implementation and computational cost reduction, the previous acoustic channel is used for the current image processing until new acoustic information is received.

6.3.5 Including the acoustic information in the SGGMM model

Having obtained the 3D location of the acoustic source, we further look at creating an acoustic channel to be used in the augmented SGGMM model. Thus, the estimated coordinates are transformed from world to image frame using the perspective projection described in Chapter 2. Additionally, we used the Tsai method for the calibration parameters of the camera as described in [151]. The obtained coordinates are used as the mean (μ) of a Gaussian model used to estimate the pixels intensities in the created acoustic channel. It has the following form:

$$f(x, \mu) = ae^{\left(\frac{x-\mu}{b}\right)^2} \quad (6.1)$$

where x represents the pixel coordinates in the image. a and b are user defined parameters that represent the scale and shape of the model.

Using this function, we attribute to each pixel x a value $I_{xs} = f(x, \mu)$ and variance $\sigma_{xs} = \sigma_0$ (set initially). Note that the latter can be used in a similar way to the parameters a and b in Equation 6.1 to determine detected region of active acoustic objects in the scene. In a such case, pixels that have been assigned higher value of σ_{xs} will have reduced chances to be accounted as foreground and vice versa).

The statistic (I_{sx}, σ_{sx}) is used for each pixel x in the image to augment the RGB vector of the SGGMM model. The latter is therefore augmented and becomes a 4D vector with the following vector and covariance matrix.

$$\mu_x = \begin{bmatrix} \mu_{c,x} & I_{xs} \end{bmatrix}, \Sigma_x = \begin{bmatrix} \Sigma_{c,x} & 0 \\ 0 & \sigma_{sx} \end{bmatrix} \quad (6.2)$$

where $\mu_{c,x}$ and $\Sigma_{c,x}$ represent the statistic of each pixel in the SGGMM model

Using the augmented vector, the new SGGMM based colour and acoustic information is used to detect the active acoustic sources in the scene.

6.3.6 Experimental tests

6.3.6.1 Evaluation of the improvement in the detection accuracy

Different tests have been done in the Unmanned Autonomous System Laboratory (UASL) to evaluate the performance of the proposed solution. Figures 6.5, 6.6 and 6.7 show selected scenes for these tests that contain moving targets. In each of these tests, one of the moving targets emits a sound pulse.

In the first scenario, presented by Figure 6.5, the energy pulse was sent from the robot in the left side in a scene containing two moving robots. In the second scenario, the sound was caused by the man as shown in Figure 6.6. In the third scenario (Figure 6.7) the sound pulse was sent from a robot in a scene containing two robots and a moving person.

The ground truth, which represents pixels belonging to the moving objects, is build according the scheme described in section 3.6.5 of chapter 3. Two types of ground truths have been built: One represents all moving target, while the second represents the acoustic moving target only.

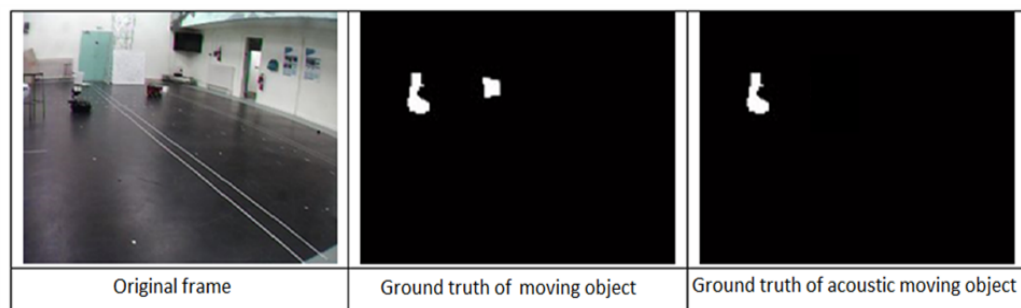


FIGURE 6.5: First test scenario with the corresponding ground truth.

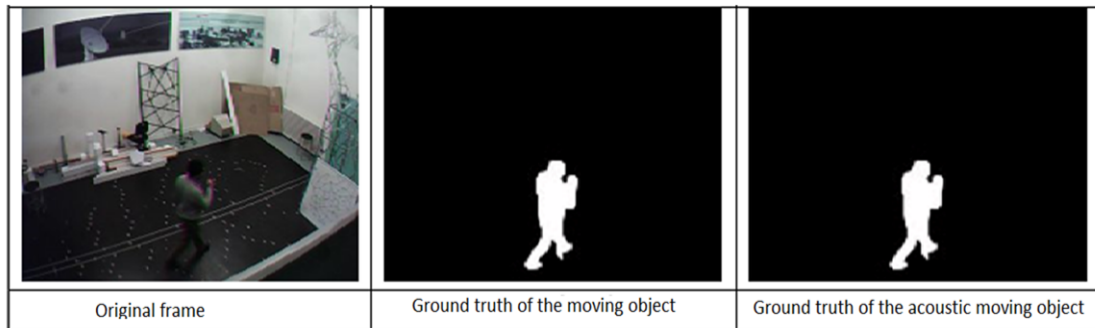


FIGURE 6.6: Second test scenario with the corresponding ground truth.

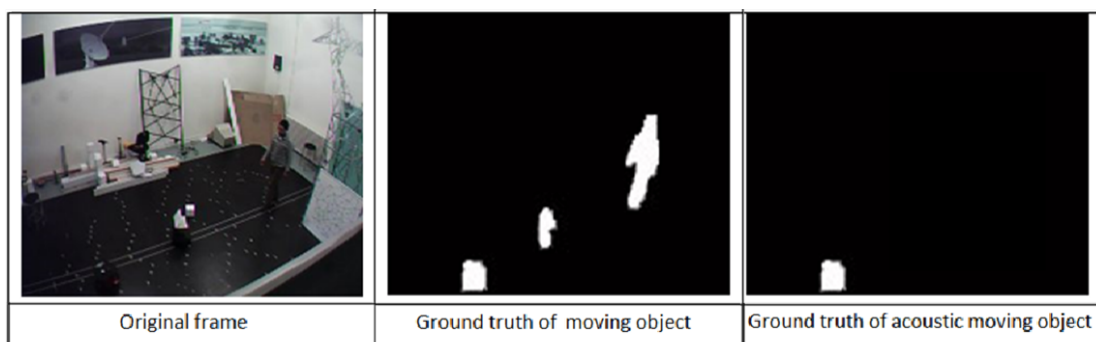


FIGURE 6.7: Third test scenario with the corresponding ground truth.

Results obtained using the augmented SGGMM fusion approach and the SGGMM based segmentation only are displayed in Figures 6.8-a and 6.8-b respectively. The ability of the fusion technique to distinguish between the active acoustic source and the mobile source is shown in the first scene, as only the moving robot that emits the sound pulse was segmented to the foreground.

In the second scene (second column in Figure 6.8), the active moving sources were hardly distinguishable from the background using the SGGMM only as the lower part of the moving target was similar to the background colour. However, by the utilisation of the proposed fusion approach, improvement in detection accuracy was achieved. This resulted in the segmentation of about the full regions of the target of interest to the foreground leading to an increase in values of the qualitative metrics of detection. These are the similarity (S) and the F-measure (F) calculated according section 3.6.2 of chapter 3.



a- Result of the SGMM based colour only



b- Result of the SGMM Augmented with the acoustic information

FIGURE 6.8: Results of SGMM colour background model and augmented SGMM with the acoustic signal for active acoustic object detection.

A similar result was obtained in the third scene (third column in Figure 6.8), in which the robot was the active acoustic source. It can be seen that SGGMM only approach failed drastically in its detection. However, with the support of the acoustic information, the full shape of the robot was captured.

Motivated by the qualitative result of detection obtained in the second scenario which include a single acoustic moving target, we evaluate the quality of detection of the two method quantitatively. Evaluation results on a sequence of 10 frames are shown in Figure 6.9. The recorded values of the similarity and the F-measure metrics clearly highlight the improvement of the detection made by including the acoustic information in the SGGMM model.

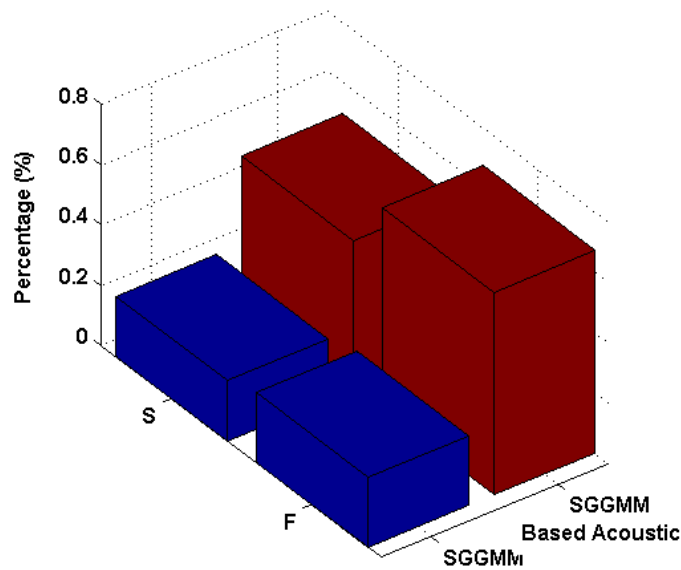


FIGURE 6.9: Quantitative evaluation of the improvement in detection

6.3.6.2 Evaluation of the improvement in localisation accuracy

Driven by the quality of results obtained from the first tests, we set a second experiment in which we track the trajectory of a mobile robot. The latter was fitted with an electronic device that emits short energy pulse at regular time intervals (0.5 sec). Figure 6.10 shows part of the setup used in the experiments. Figure

6.11 shows the path followed by the robot with the corresponding measurements obtained from the camera, the acoustic sensors and the proposed fusion approach.

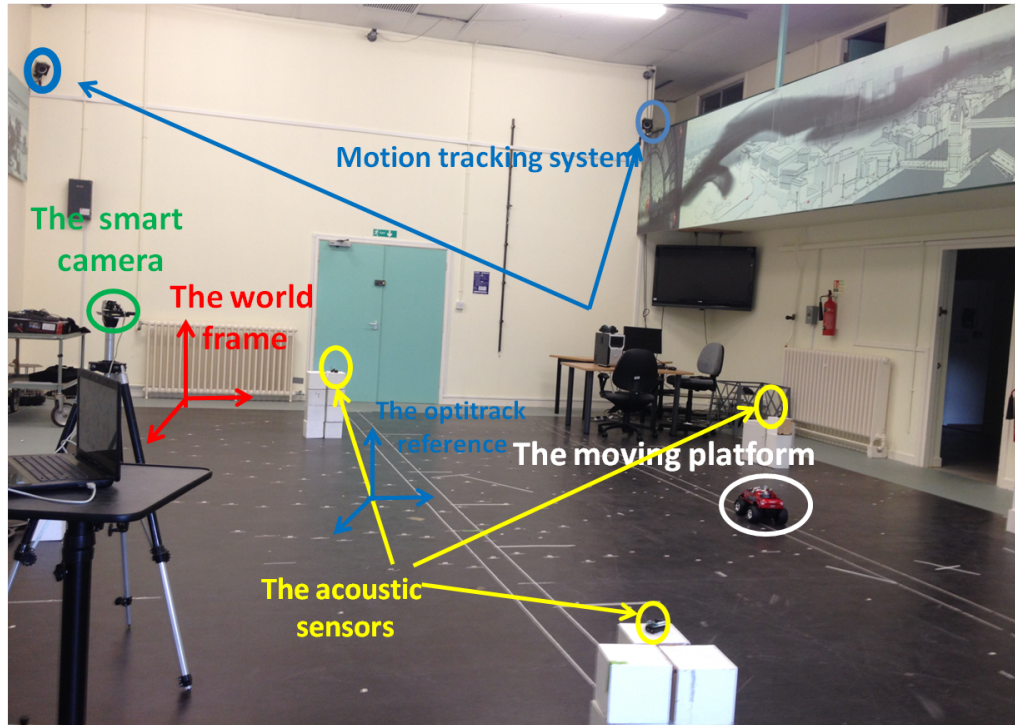


FIGURE 6.10: Setup used for the robot localisation using the proposed fusion approach

To compare the performances of the proposed fusion approach with results obtained using the acoustic measurements only. We use the average distance of error (DE) in localisation per estimate, given by

$$DE = \frac{\sum \sqrt{(X_{GT} - X_{Est})^2 + (Y_{GT} - Y_{Est})^2}}{n} \quad (6.3)$$

where (X_{Gt}, Y_{Gt}) represents the ground truth coordinate obtained using the motion tracking system installed in the UASL laboratory. (X_{Est}, Y_{Est}) represents the estimated coordinates, while $n = 200$ is the number of measurements.

Figure 6.12 shows the results obtained using the fusion approach compared to using separate types of Data. It clearly shows the accuracy improvement of the active acoustic sources localisation using the proposed fusion method over

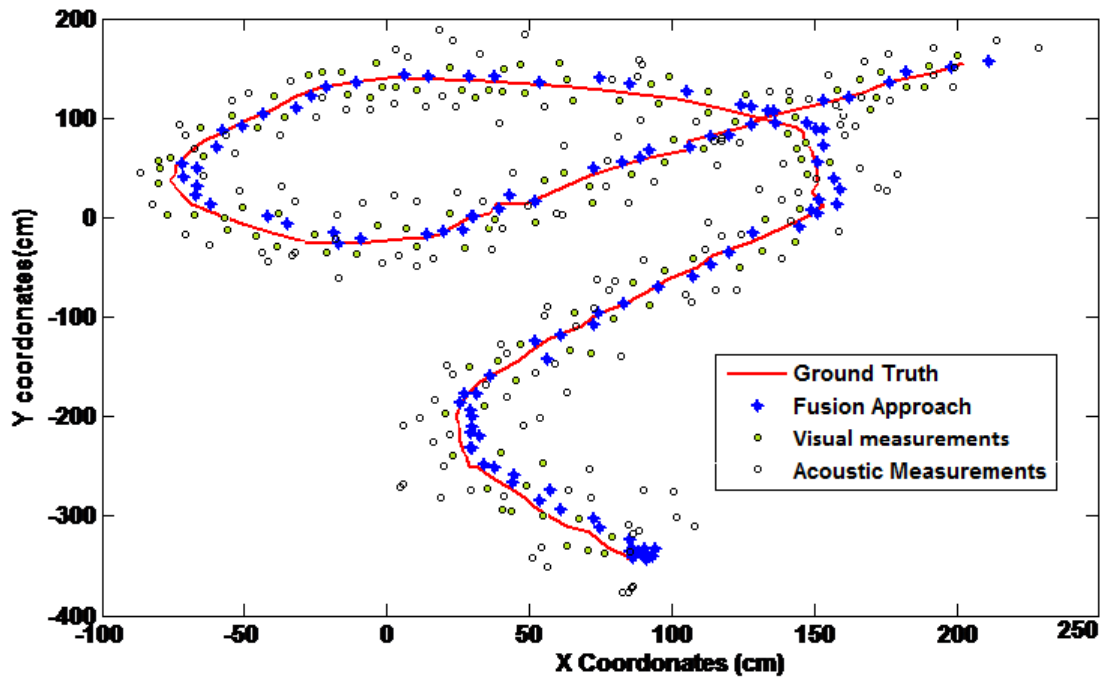


FIGURE 6.11: Moving platform trajectory with the collected measurements

measurements obtained using the acoustic and the visual data when these measurements are used in a separate manner.

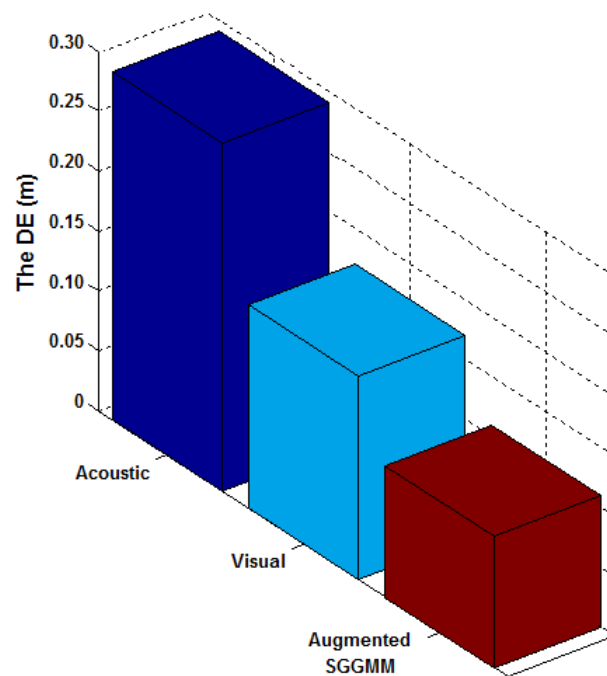


FIGURE 6.12: Accuracy Comparison between localisation based on acoustic measurements only and SGMM based acoustics

6.4 Second part: Cooperative localisation and tracking

One of the main activities of a surveillance system is to ensure the tracking of specific targets of interest with high accuracy. Designing an architecture that combines measurements of the different sensors available is a well investigated subject. The centralised/decentralised architecture fusion based on Kalman filtering [144, 152–157] was one of the most investigated solutions to this problem. Such a solution shows a high ability in dealing robustly with problems of estimation under noisy measurements. In this section, we investigate the usage of this approach for the problem of tracking using a distributed heterogeneous sensors network featured by a limited communication bandwidth. Through experimental tests, we evaluate the accuracy of tracking obtained by the combination of two different modalities of measurements. These modalities are the acoustic and the video. The evaluation involves performance comparison between the information fusion (IF) [158] and covariance intersection scheme (CI)[144, 145] and to the accuracy achieved using single sensor modality.

In what follows, we firstly present the centralised/decentralised architecture of fusion in section 6.4.1. Having presented the CI algorithm in Section 4.4.2 4, a presentation of the IF will be given in section 6.4.2. The motion models studied are given in section 6.4.3, while in section 6.4.4 we describe the tracking scheme at each of the local tracking level. Description of the experiments completed and the experiment results are given in section 6.4.6 and 6.4.7 respectively. Finally we summarise the overall finding at the end of the chapter.

6.4.1 Fusion based on centralised/decentralised architecture

A fusion system based on the centralised/decentralised architecture starts by estimating different state vectors and their associated covariance matrices for each local level. These are passed to the central fusion level for high level tracking. In case of multiple targets tracking, the track fusion centre performs a track association process to determine the sensor level track that correspond to the true target [153, 159]. Having determined the likely association, the fusion centre collects the state vectors and predicts each one in a synchronous fashion. Adopting this technique to combine the two data modalities requires using a recursive estimator for each measurements type. For simplicity, the same filtering strategy is used for both modalities (acoustic and video). This is achieved using the unscented Kalman filter (UKF). Figure 6.13 depicts the main components of this architecture. It shows briefly the various hardware and software tools used to fuse these two different types of data. In this architecture the same state vector is adopted at the two tracking levels, it is given by the following:

$$X_{i,k+1} = \begin{bmatrix} x_{i,k+1} \\ \dot{x}_{i,k+1} \\ y_{i,k+1} \\ \dot{y}_{i,k+1} \\ \varphi_{i,k+1} \end{bmatrix} \quad (6.4)$$

where $x_{i,k+1}, y_{i,k+1}$ represents the Cartesian coordinates of the tracked target; $\dot{x}_{i,k+1}, \dot{y}_{i,k+1}$ represent the projection of the speed on the X and Y axis. The variable $\varphi_{i,k+1}$ is the target heading. The variable i ($i = a$ for acoustic and v for the video) represents the type of the measurements source. The state and measurement vectors at local tracking levels $(i)_{i=a,v}$ are both transformed to a

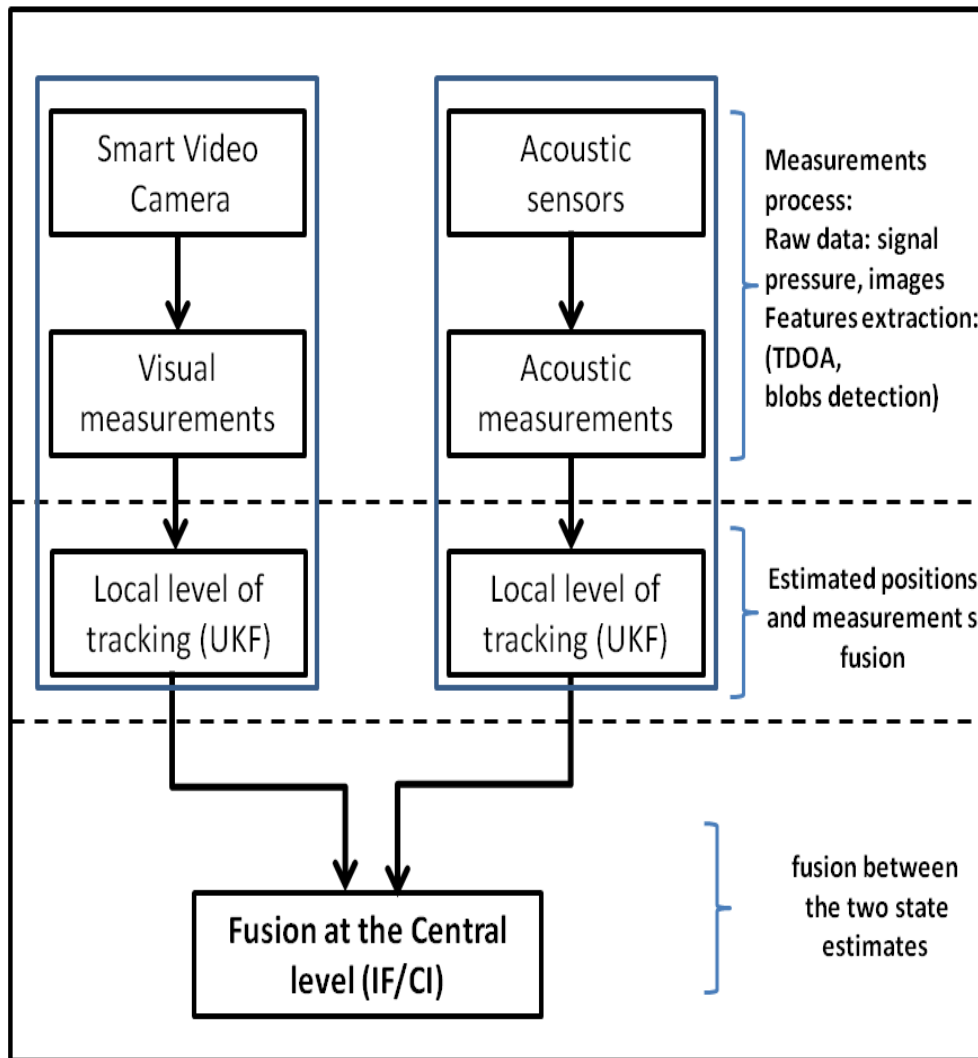


FIGURE 6.13: Main component of the centralised/decentralised architecture of fusion.

unique Cartesian reference. The object state follows a discrete time dynamic model as given in Equation (6.5):

$$X_{i,k|k-1} = F(X_{i,k|k}, w_i) \quad (6.5)$$

where w_i is the process noise, assumed to follow a Gaussian distribution of zero mean and covariance matrix Q_i . The observation follows the measurement equation given by:

$$Z_{i,k} = H^{(i)}(X_{i,k}, v_i); \quad (6.6)$$

$Z_{i,k}$ is the measurement vector at a local tracking level i at time k , $v_{i,k}$ represents the measurement noise vector characterised by the noise covariance matrix R_i . The predicted state vectors and covariance matrices at local fusion levels are combined to produce a composite state and covariance matrix for each track using the fused state.

6.4.2 Fusion at the central level

In the presented centralised/decentralised fusion we evaluated the performance of two classical algorithms (Figure 6.13), the first is the CI algorithm presented in chapter 4, while the second is the IF which is given by the following equations:

The inverse of the combined covariance $P_{f,k|k}$ at instant k , it is given by:

$$P_{f,k|k}^{-1} = P_{f,k-1|k-1}^{-1} + (P_{a,k|k}^{-1} - P_{a,k|k-1}^{-1}) + (P_{v,k|k}^{-1} - P_{v,k|k-1}^{-1}) \quad (6.7)$$

The state estimate vector which is given by:

$$\begin{aligned} P_{f,k|k}^{-1} \hat{X}_{k|k} &= P_{f,k|k-1}^{-1} \hat{X}_{k|k-1} + P_{a,k|k}^{-1} \hat{X}_{a,k|k} - P_{a,k|k-1}^{-1} \hat{X}_{a,k|k-1} \\ &\quad + P_{v,k|k}^{-1} \hat{X}_{v,k|k} - P_{v,k|k-1}^{-1} \hat{X}_{v,k|k-1} \end{aligned} \quad (6.8)$$

which can be reformulated to:

$$\begin{aligned} \hat{X}_{k|k} &= P_{f,k|k} [P_{f,k|k-1}^{-1} \hat{X}_{k|k-1} + P_{a,k|k}^{-1} \hat{X}_{a,k|k} - P_{a,k|k-1}^{-1} \hat{X}_{a,k|k-1} \\ &\quad + P_{v,k|k}^{-1} \hat{X}_{v,k|k} - P_{v,k|k-1}^{-1} \hat{X}_{v,k|k-1}] \end{aligned} \quad (6.9)$$

An alternative approach to the IF is to use the covariance matrix obtained at the central level to update the local level known as fusion with feedback [152]. However, this approach is avoided because of the high communication cost it induces.

6.4.3 The motion model

Modelling the dynamic motion of a target is a necessary step to perform a tracking using any type of Kalman based filtering method. Most ground targets (pedestrian, animals, cars,... etc) take one or a mixture of motion models. These models can either be stationary, moving with constant speed, turning with rate, moving with acceleration or turning with accelerating turn rate. The performance of the centralised/decentralised fusion approach is assessed using the first three motion models. For the static case, the transition matrix F introduced in Equation (6.5) is given as:

$$F = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (6.10)$$

For a linear motion model with a constant velocity, F is written as:

$$F = \begin{bmatrix} 1 & \Delta t & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (6.11)$$

with Δt is the sampling time.

In the case of a transition model with a known turning rate the motion function is given by the following:

$$F = \begin{bmatrix} 1 & 0 & \frac{\sin(w\Delta t)}{w} & 0 & -\frac{1 - \cos(w\Delta t)}{w} & 0 \\ 0 & \cos(w\Delta t) & 0 & -\sin(w\Delta t) & 0 & 1 \\ 0 & -\frac{1 - \cos(w\Delta t)}{w} & 1 & \frac{\sin(w\Delta t)}{w} & 0 & 0 \\ 0 & \sin(w\Delta t) & 0 & 0 & \cos(w\Delta t) & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (6.12)$$

with w is the turn rate.

6.4.4 Local level of tracking

In this section, we show how each local level of tracking, presented in Figure 6.13, works to deliver the estimate of the state vector with the corresponding covariance to the higher level of tracking.

6.4.4.1 Local level tracking using the acoustic data

Robust sound source localisation with uncertainty estimation is needed as this will be employed in the proposed fusion scheme with visual data. In this context, only few solutions are reported in the literature to adopt a recursive approach for the acoustic source localisation. A solution based on the extended Kalman filter (EKF) proposed in [117] has been shown to deal reasonably well with this problem [117]. However, due to the non-linear nature of the measurement function in an acoustic localisation problem based on TDOA, we propose the use of the UKF instead of the EKF. This filter has been shown to be more practical for systems with highly non-linear dynamic or measurement functions. One of the advantages of this technique over the EKF is that it does not require a direct calculation of the Jacobians, nor the Hessians for the algorithm implementation. Moreover,

the overall number of computations has nearly the same order as the EKF [160]. An in-depth comparison between the two algorithms showing better estimation accuracy of the UKF when it is used for simultaneous localisation and mapping (SLAM) problem can be found in [161].

6.4.4.2 Problem modelling

To model the problem of acoustic localisation using a recursive filtering technique, the state space utilised to model the acoustic object motion is the same given in Equation 6.5. The state update model is given by Equation 6.4.

From the acoustic signal model equation using TDOA localisation based approach, the observation matrix can be written as:

$$H = \begin{bmatrix} TDOA_1 \\ \vdots \\ TDOA_{n-1} \end{bmatrix} \quad (6.13)$$

with n represents the number of TDOA measurements. By substitution of the matrix terms with the corresponding Euclidean distance we get:

$$H = \frac{1}{s} \begin{bmatrix} \sqrt{(x_k - x_1)^2 + (y_k - y_1)^2} - \sqrt{(x_k - x_2)^2 + (y_k - y_2)^2} \\ \vdots \\ \sqrt{(x_k - x_1)^2 + (y_k - y_1)^2} - \sqrt{(x_k - x_4)^2 + (y_k - y_4)^2} \end{bmatrix} \quad (6.14)$$

with (x_k, y_k) the coordinates of the target to be estimated, and are given in state vector (Equation 6.4) and $s = 343m/s$ represents the sound speed, $[(x_1, y_1); (x_2, y_2)]$ the position of the first pair of microphones while $[(x_1, y_1); (x_4, y_4)]$ the position of the third pair of microphones. The development of the state estimate in the UKF is specified using a minimal set of selected sample points obtained from

the unscented transformation (UT). The selected points capture the true mean and covariance of the process, and when propagated through the true non-linear system, it captures the posterior mean and covariance up to the 3^{rd} order (Taylor series expansion) for any non-linearity [160].

6.4.4.3 The Unscented transformation (UT)

The unscented transformation (UT) [160] is adopted by the UKF to calculate the statistics of a random variable that undergoes a non-linear transfer function. For a random variable x (with dimension n , a mean \bar{x} and covariance P_x) that is propagating through a non-linear function f (with $y = f(x)$), the statistics computation of y requires forming a matrix \mathcal{X} of $2 \times n + 1$ Sigma vector χ_i with the corresponding weights \mathcal{W}_i . This is completed through the following set of equations:

$$\left\{ \begin{array}{l} \chi_0 = \bar{x} \\ \chi_i = \bar{x} + (\sqrt{(n+\lambda)}P_x)_i, i = 1, \dots, n \\ \chi_i = \bar{x} - (\sqrt{(n+\lambda)}P_x)_{i-n}, i = n+1, \dots, 2n \\ \mathcal{W}_0^m = \frac{\lambda}{n+\lambda} \\ \mathcal{W}_0^c = \frac{\lambda}{n+\lambda} + (1 - \alpha^2 + \beta) \\ \mathcal{W}_i^m = \mathcal{W}_i^c = \frac{1}{2(n+\lambda)} \end{array} \right. \quad (6.15)$$

with $\lambda = \alpha^2(n+k) - n$ is a scaling parameter, α determines the spread of the Sigma points around x and it is usually set to a small positive value. k is a secondary scaling parameter, while β is used to incorporate prior knowledge of the distribution (for Gaussian distributions $\beta = 2$ is optimal). $(\sqrt{(n+\lambda)}P_x)_i$ is the i^{th} row of the matrix square root.

Figure 6.14 shows the propagation of these Sigma vectors through the function f as follows:

$$\mathcal{Y}_i = f(\mathcal{X}_i), i = 0, \dots, 2n. \quad (6.16)$$

while the mean and covariance for y are approximated using a weighted sample mean and covariance of the posterior sigma points,

$$\bar{y} = \sum_{i=0}^{2n} \mathcal{W}_i^m \mathcal{Y}_i \quad (6.17)$$

$$P_y = \sum_{i=0}^{2n} \mathcal{W}_i^c \{\mathcal{Y}_i - \bar{y}\} \{\mathcal{Y}_i - \bar{y}\}^T \quad (6.18)$$

The approach of the UT results in approximations that are claimed to be accurate to the third order for inputs of non-linear systems [160].

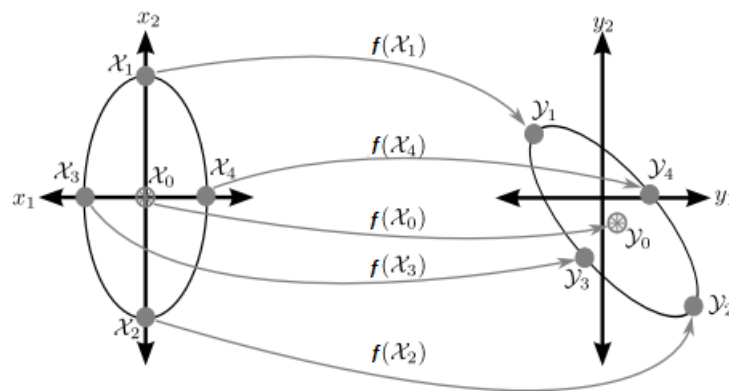


FIGURE 6.14: Illustration of the unscented transform(UT)

6.4.4.4 The UKF Algorithm

The UKF is a direct extension of the UT to the recursive estimation in Equation (6.15). The UT sigma point selection scheme is applied to the new augmented state $X_k^a = [X_k^T v_k^T n_k^T]^T$ obtained by the concatenation of the original state and noise variables. Typically in this filter, the two phases (prediction and update) alternate, with the prediction advancing the state until observation is provided. The basic equations of the UKF are given in Figure 6.15.

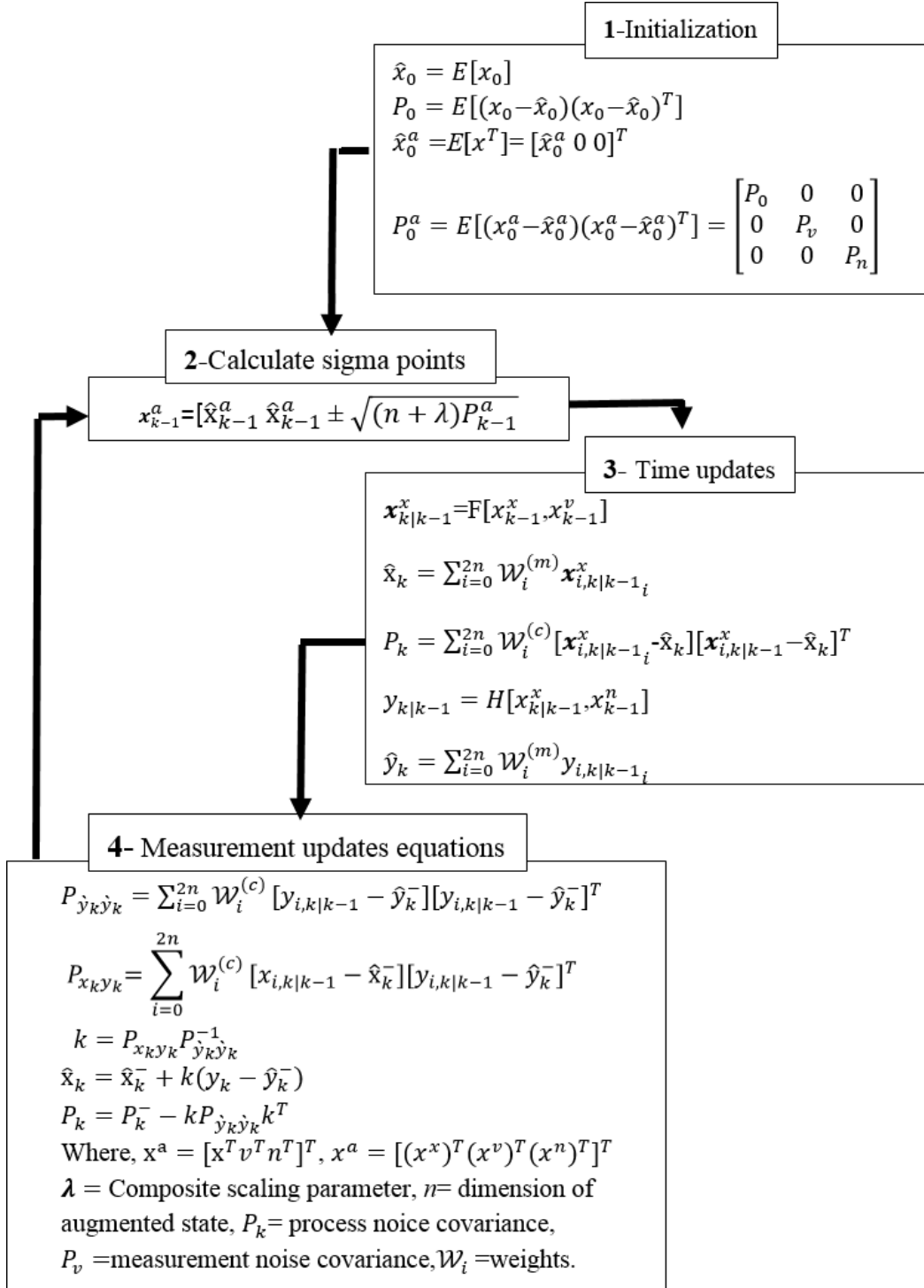


FIGURE 6.15: The alternation between the UKF steps

6.4.4.5 Results and discussions

Representative results of using the UKF to estimate the location of an acoustic source are shown in Figure 6.16. The true location of the source is the coordinate (4 m, 12 m) in a Cartesian reference system. Figure 6.16-a represents the estimation of the x-coordinate with regards to the true values, while the development of its corresponding covariance parameter is shown in figure 6.16-b. Figures 6.16-c and 6.16-d show the development of the y-coordinate estimation in comparison to the true value of the position and the variation of its corresponding covariance parameter. The new obtained results show how well the UKF performs to

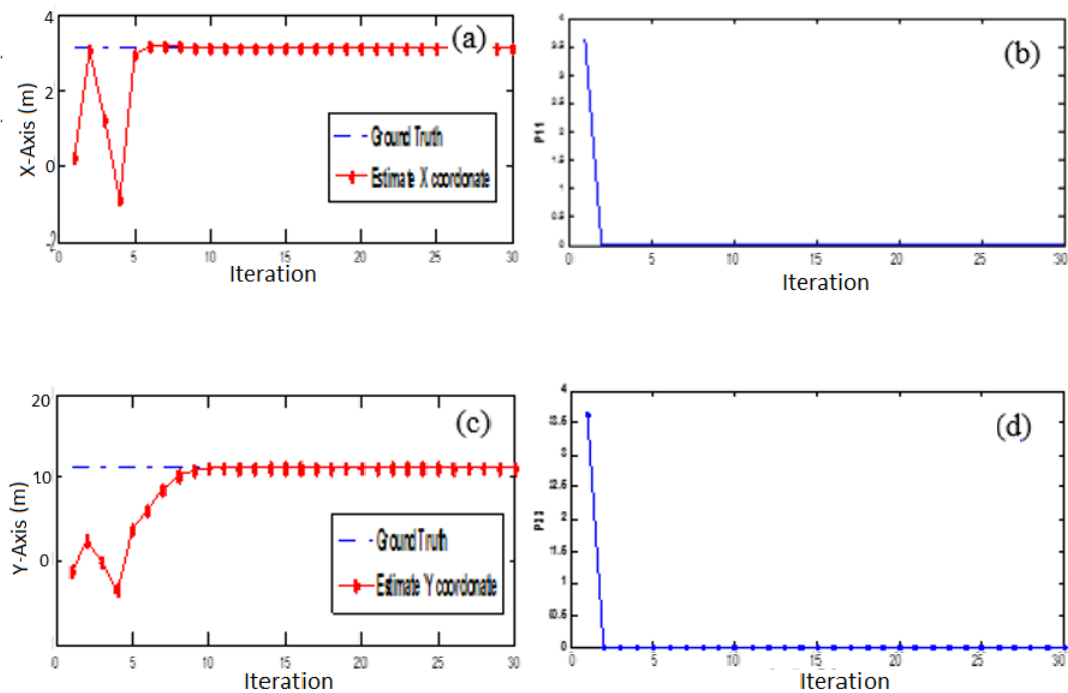


FIGURE 6.16: Behaviour of the UKF in reaching the optimum solution for the sound source localisation

estimate the source location. It converges right quickly to the position of the acoustic source. The UKF started to perform well after just a reduced number of iterations (10). The error margin in errors between the estimation and the measurement is very small (21 cm for the x-axis and 23 cm and for the y-axis, respectively).

6.4.4.6 Local level of tracking using the visual Data

In this section, we briefly discuss the main steps to transform a location of detected moving object to a visual measurement. Figure 6.17 depicts the steps of this operation. Once a moving object is detected using a foreground/background

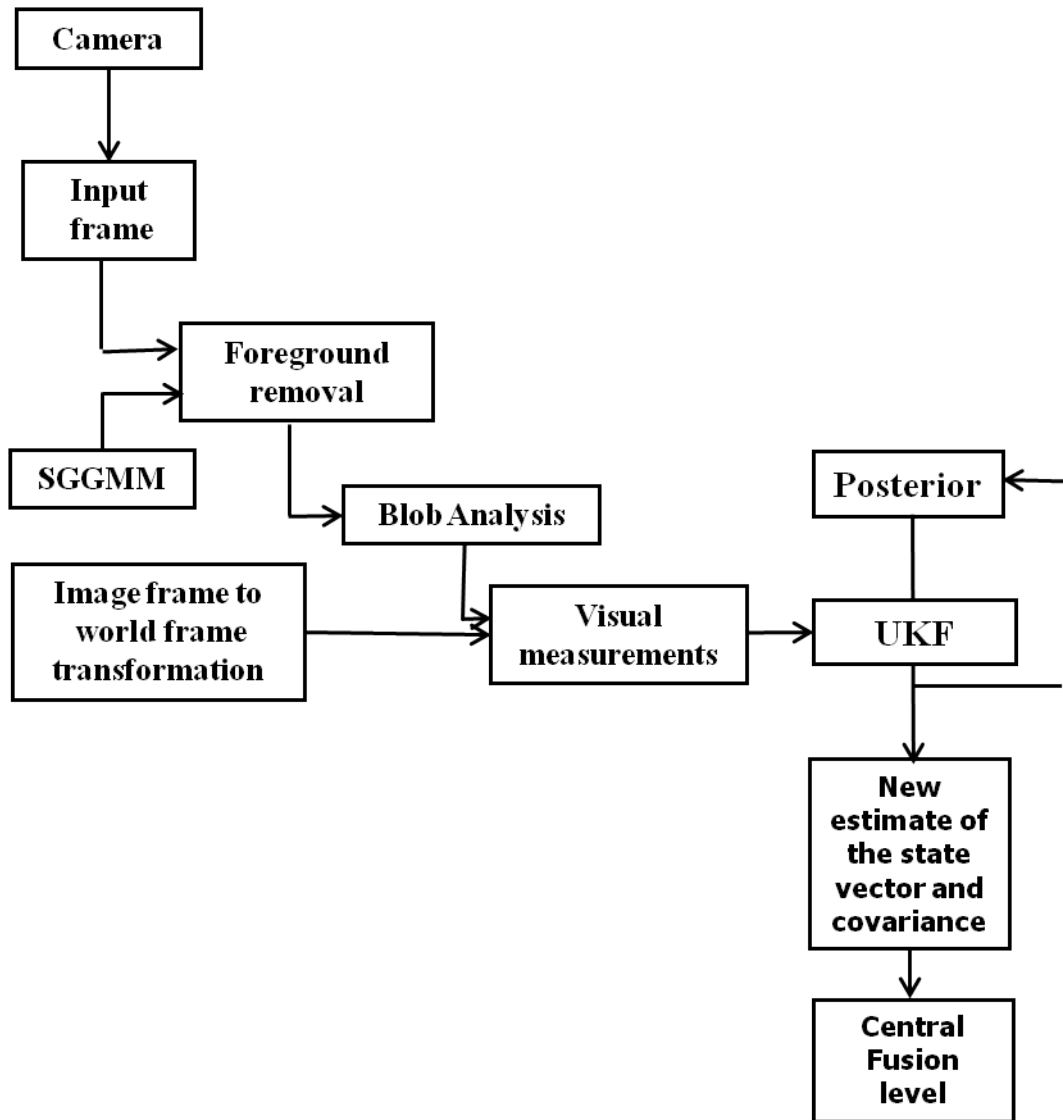


FIGURE 6.17: UKF feeding with visual measurements.

method (we used the SGMM for this step), the coordinates of the centroid of the detected moving blob is obtained through a transformation from image frame to world frame. The same world reference and the same state model are used

for the acoustic measurement is adopted to register the estimated position of the moving objects based on the visual data.

The measurement matrix $H^{(v)}$ we use here is the following:

$$H^{(v)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad (6.19)$$

This function ensures a direct reading of the projected centroid corresponding to visual appearance of the detected moving object.

6.4.5 Experimental setup

Feasibility and performance of the proposed fusion approach in distributed sensor networks frame-work was tested experimentally. In the following experiments, a distributed hierarchical tracking system is built as shown in Figure 6.18 where the lowest level is constituted by heterogeneous sensors (a CITRIC camera and four Micaz motes with the MTS310 sensors board). The highest level is the fusion centre composed of a desktop used to collect and process the provided measurements. Both types of sensors are set to detect and send measurements corresponding to a mobile platform. The latter was fitted with an electronic device to emit a burst of sound (a sinusoidal burst of 4 kHz which lasts for 0.2 sec). The projection from 2D to 3D was done using the method proposed in [162], while the acoustic sensors was placed at known position. The performance evaluation has been done at different acquisition rates: the first was 0.5 sec for the acoustic and 0.3 sec for the video. For the second, the video measurements acquisition rate 0.5 sec, while it was equal to 0.5 sec for the acoustic.

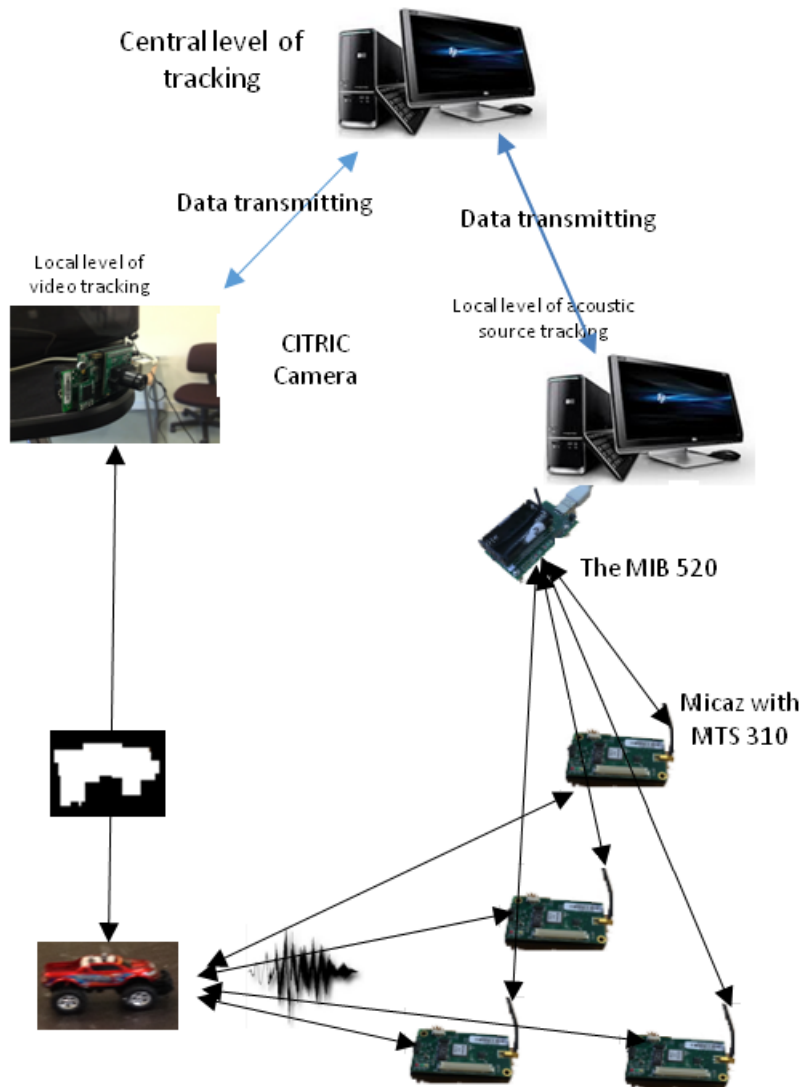


FIGURE 6.18: Setup used for the centralised/decentralised architecture of fusion implemented

6.4.6 Experiments

The fusion scheme using the method proposed previously was tested in three different scenarios. Each scenario presents one form of the motion models presented in section 6.4.3. The first scenario is for an object of fixed position with simple movements. These movements are centred around a given coordinate in relation to the reference system. The coordinates (4 m, 4 m) have been chosen as the centre of motion. Results of the first scenario are shown in Figure 6.19-a. In the second experiment, we evaluate the performance of the tracking when the target follows a linear motion model. This is shown in Figure 6.19-b. In the third experiment, the performance of the fusion approach is tested when the target moves in a circular trajectory with a fixed turn rate. The corresponding trajectory with the different estimation results are given in Figure 6.19-c.

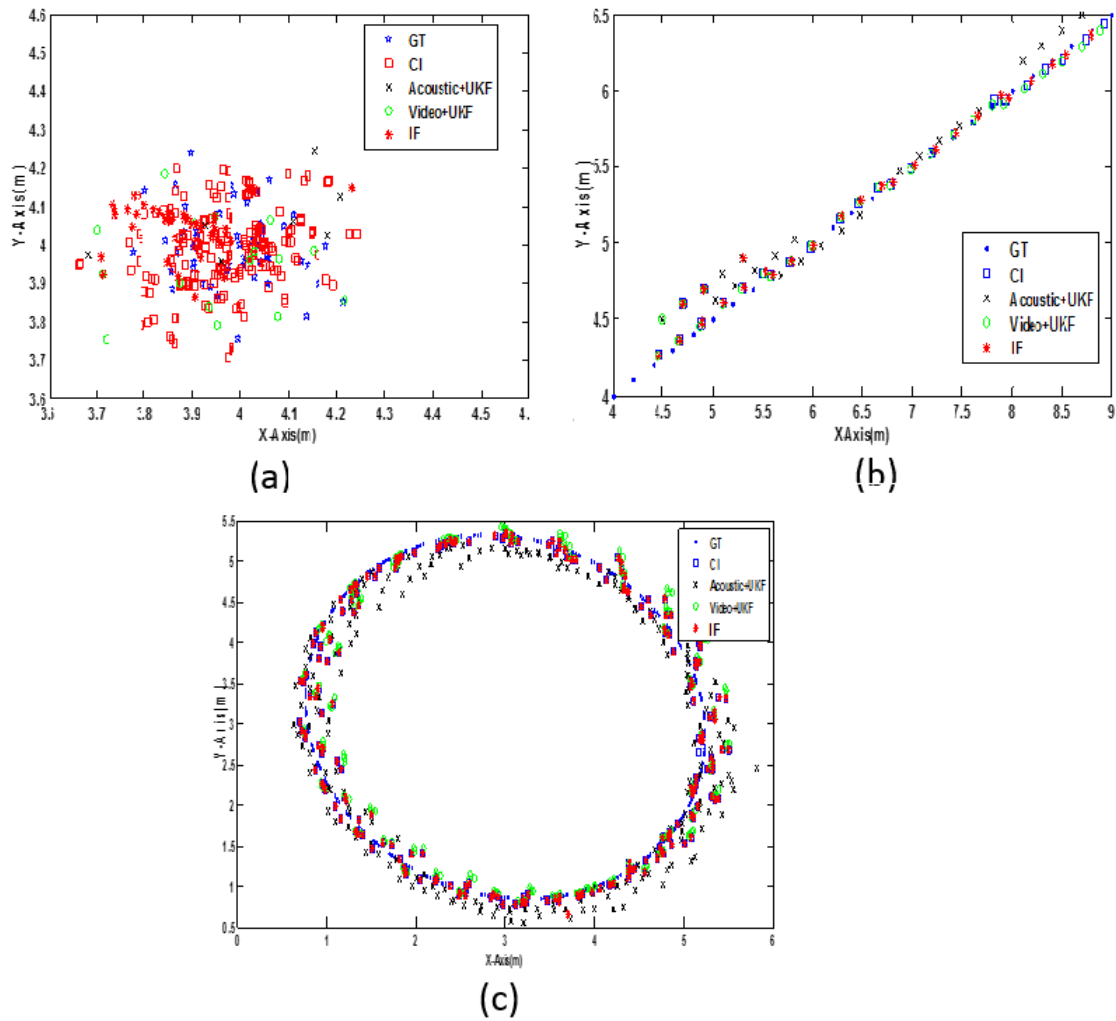


FIGURE 6.19: Results obtained from the tracking scheme applied to the different motion models: a:Stationary, b:Linear motion, c:Circular trajectory

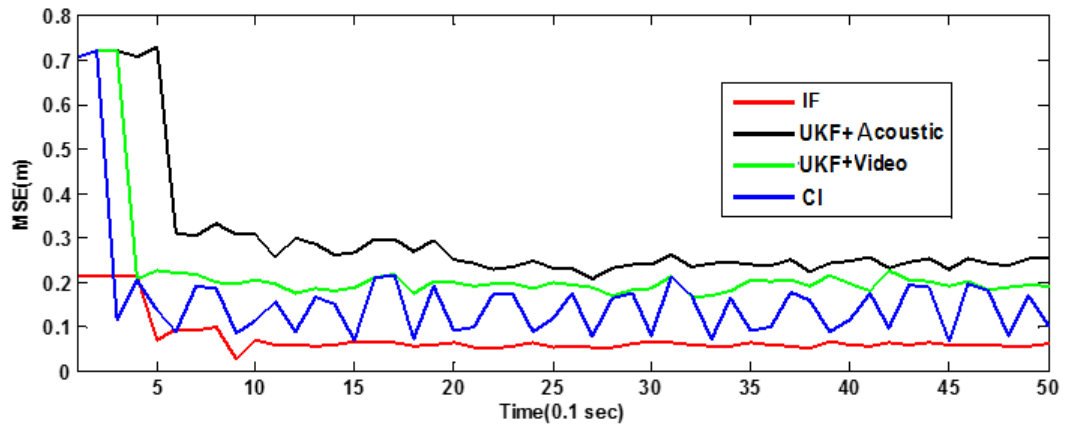
6.4.7 Evaluation result

In the following, the accuracy evaluation of the proposed tracking scheme applied to the different motion models is discussed. The metric used for the evaluation is the mean square of errors (MSE) estimated for a number of 30 measurements for each experiment, while the ground truth is measured using a motion tracking system.

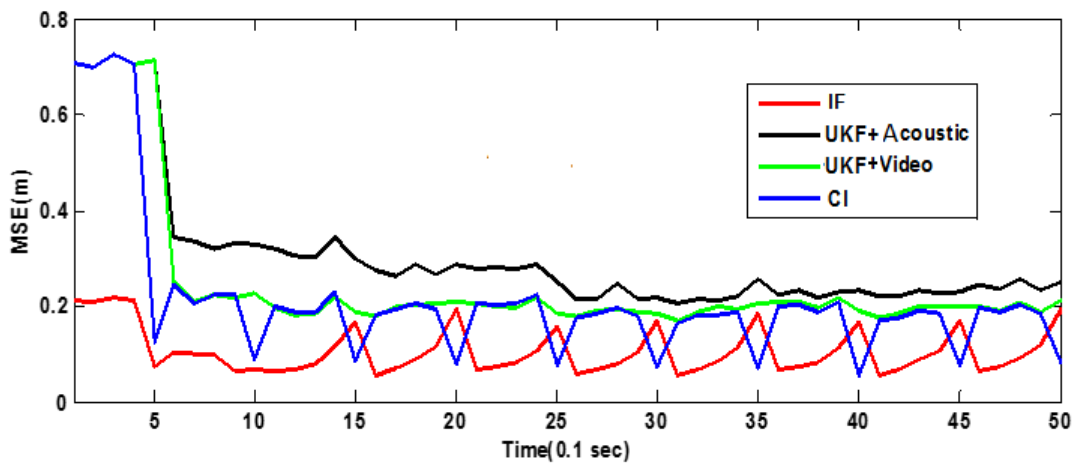
6.4.7.1 For a stationary acoustic source

To test the performance of the centralised/decentralised architecture with the information fusion(IF) and the covariance intersection (CI), and to compare their performance to a single type of data (either the visual or the acoustics) for the stationary case, two experiments have been conducted using the two different acquisition rates as explained in section 6.4.5.

Figure 6.20 shows the obtained results. It can be seen that the IF which includes the information corresponding to the previous state estimate, outperforms the CI method. This result can be noticed specially in the situation where the acquisition rate for both types of measurements is not the same (Figure 6.20-a). Similar performance of the evaluated fusion methods can be observed when both measurements (video and acoustic) are acquired at similar rate (0.5 sec) as shown in Figure 6.20-b. However, an improvement in the accuracy of the CI in moments of measurements acquisition is noticed making this fusion method most suitable for such cases. Hence, higher accuracy in the tracking using the CI is expected for scenarios where the measurements are provided with higher acquisition rates.



a- Estimation obtained at variable acquisition rate of measurements



b- Estimation obtained at similar acquisition rate of measurements

FIGURE 6.20: MSE obtained using different algorithm for static object position estimation

6.4.7.2 Moving in a linear path

For a linear path scenario, a high level of accuracy can be noticed in the performance of CI over the IF. It also gives better results using than the obtained using a single sensor (Figure 6.21). This observation is valid for the two different case of measurements acquisition rates. An improvement to the accuracy of the centralised/decentralised architecture based on the IF is noticed at moments of acoustic measurements acquisition. However, in such case, the accuracy of this fusion method can only reach the one of the CI.

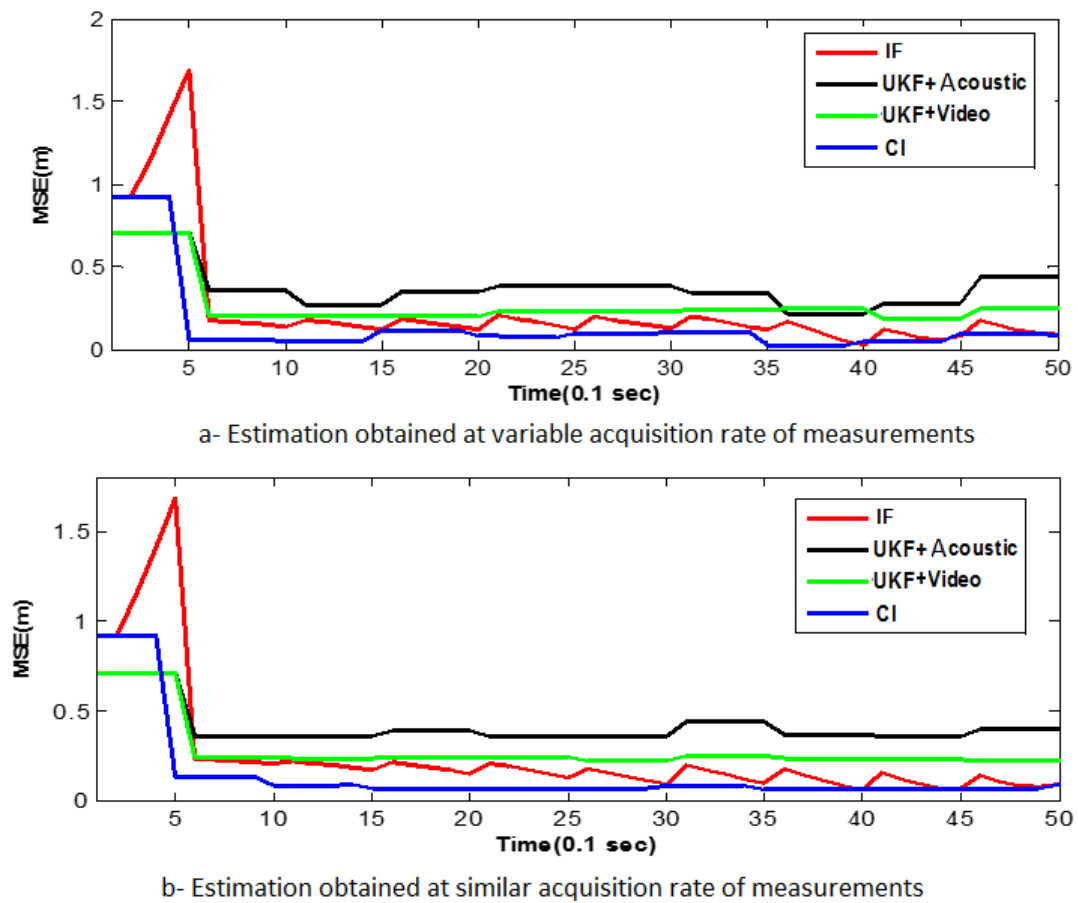


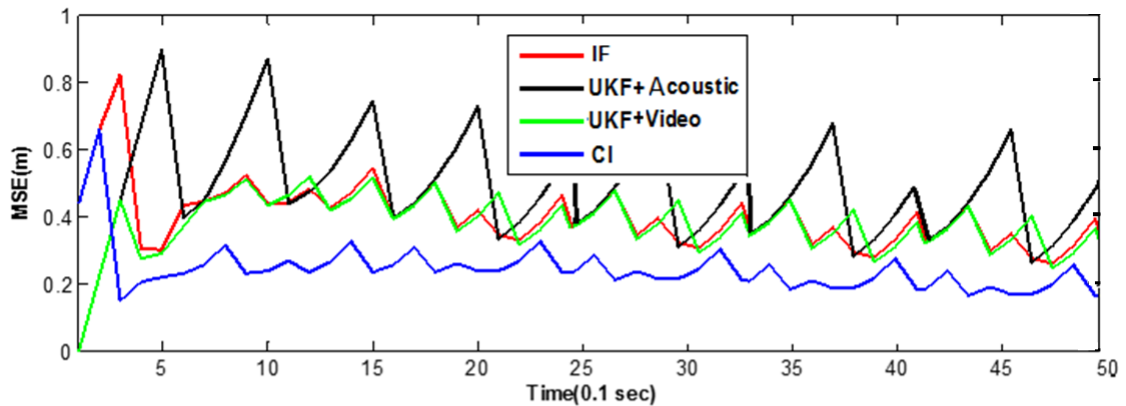
FIGURE 6.21: MSE obtained using different algorithm for tracking the active acoustic object following linear trajectory

6.4.7.3 Motion with constant turn rate

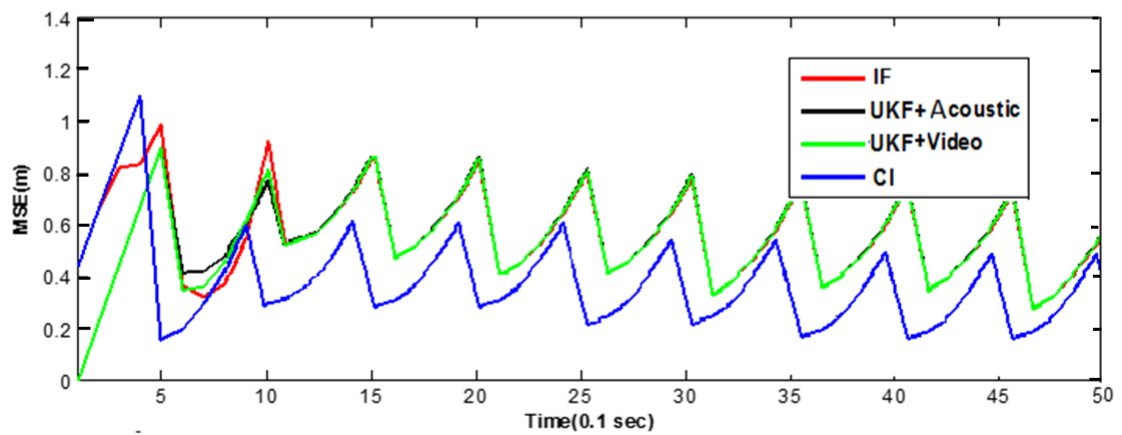
Similarly to the first two scenarios, experiments using two different acquisition rates have been conducted for the constant turn rate scenario. The results of the first experiment (Figure 6.22 a) corresponding to a variable acquisition rate highlights the performance of CI over the IF. Consequently, the CI also delivers better performance than using one single type of data.

When the two different measurements (video and the acoustics) are received at similar and relatively lower acquisition rate (0.5 sec) (Figure 6.22-b), we can notice the degradation in the performance of the tracking at the local tracking level corresponding to the video modality compared to the first experiment. The

Accuracy of the latter, become similar to that one of the acoustics which is caused mainly by the errors in the state estimation process that is expected to be higher in a such dynamic model. We can notice also that although there is degradation in the results obtained the CI, these results remains the most accurate.



a- Estimation obtained at variable acquisition rate of measurements



b- Estimation obtained at similar acquisition rate of measurements

FIGURE 6.22: MSE recorded using the different algorithms for tracking the active acoustic object following circular trajectory

6.5 Conclusion

This chapter covered two fusion methods of acoustic and visual data in a distributed sensor networks. The first concerns augmenting the 3D RGB vector of the SGGMM model with the acoustic signal. The method has shown its ability to highlight the presence of the active acoustic source in the scene using a visual tool. This led to reaching higher accuracy in the localisation provided that camera is properly calibrated.

Results from the experiments carried out have demonstrated that the proposed technique allows a significant improvement in detecting the active acoustic sources. Additionally, the method showed an overall accuracy improvement in estimating the position of the acoustic sources when the visual information is included.

For the second method of fusion, we investigated the use of the centralised/decentralised architecture of fusion of the acoustic and the visual data in distributed sensors networks. Evaluation results show higher accuracy of this fusion architecture compared to using one single type of data. Additionally, higher accuracy can be achieved using the CI than by using the IF though the little information it uses for the estimation.

This page is intendedly left blank

Chapter 7

Conclusion and Future Work

This thesis has studied various techniques targeting the robustness and applicability of detection and localisation in distributed sensor networks. After thorough analysis and investigations, solutions have been provided where it was thought necessary and challenging.

In chapter 3, the problem of visual detection was investigated by introducing a cost efficient visual change detection method. This method adopted a spatial global Gaussian mixture models (SGGMM) to model the background based on RGB colour. The proposed method showed high accuracy in detection of moving objects in image sequences in favourable conditions. For challenging conditions caused by sudden changes in luminosity, a combination of pixel uncertainties with colour in the SGGMM model was approached to deal efficiently with problem related to background motion within the scenes. The evaluation of the proposed method has demonstrated the accuracy in detection and the suitability of its implementation in embedded camera sensor network nodes, which present reduced computation capabilities.

In future work, further accuracy evaluation in handling problems relating to shadow and camera jitter (due to severe background variation) is needed. Also,

due to the low computational cost in embedded systems, further investigation is required for using the proposed background subtraction method in a collaborative scheme of detection and tracking.

In chapter 4, the problem of detection of moving objects from a moving camera has been investigated. We have shown that by using optical flow based tracking better results can be achieved for motion detection in such challenging conditions. To guarantee respect of constancy on the intensity constraint under which the optical flow should work, we proposed the use of a robust image registration method. The latter, adopted the H_∞ filter which takes into account feature uncertainties in image registration.

Despite the efficient tools used to counter the miss registration problem, the overall estimated optical flow is shown to be corrupted with noise. Hence, we proposed a solution based on the spatial Gaussian Mixture Models. The latter has shown its ability to deal more efficiently with the investigated problem. For future work, further investigation is needed for adopting this approach in surveillance systems based on PTZ cameras;

In chapter 5, we proposed adapting a trust region based method to deal efficiently with the optimisation problem of the acoustic source localisation in WSN. Through experimental evaluations, we showed the efficiency and the accuracy of the proposed approach in comparison to a linear search based technique. The Double Dogleg method investigated in this chapter combines the advantages of the steepest-descent method, which is robust and numerically stable for initialisation far from the solution and the Gauss-Newton technique, which is featured by a rapid convergence toward the aimed solution.

An experimental formulation of the acoustics-induced uncertainties and their magnitude was also covered in this chapter, to reach a solution with the highest accuracy using the provided measurements.

In Chapter 6, solution of an architecture of fusion for active acoustic source localisation in a heterogeneous sensors network was the main focus. The first proposed solution relied on the trust region double dog leg for the acoustic source localisation, while the Spatial Global Gaussian Mixture Model (SGGMM) was used for combining the estimated acoustic source location with the vision detection model. The experimental results demonstrated the solution feasibility. Moreover, this fusion approach allowed important improvement in detection and localisation accuracy of targets of interest.

For future works, we suggest the use of advanced acoustic sensing devices with high signal processing capabilities. This will enable the technique to deal efficiently with complex scenarios. These can involve scenarios related to security enhancement in public spaces such as in the case of aggression detection or tracking targets of special acoustic features. The second proposed solution in this chapter, was to cover the basic technical issues related to fusion between the obtained measurements of two heterogeneous data modalities. To this aim, the performance of a centralised-decentralised fusion was evaluated using both the covariance intersection and the information fusion scheme.

Experimental results has shown that overall, this approach of fusion, for distributed networks, is able to provide higher accuracy in the localisation and tracking than by using one single type of measurement only. We have also shown the cost efficiency of a fusion scheme based on the covariance intersection compared to the information fusion. For the tracking, we covered principal dynamic models of active acoustic targets (stationary, linear motion, turning). In case of targets manoeuvring, the interacting multiple model (IMM) can be used.

This page is intendedly left blank

Bibliography

- [1] H. Opower, *Multiple view geometry in computer vision*, second edi ed., Cambridge University Press, Ed., 2002, vol. 37, no. 1.
- [2] B. Cyganek and J. P. Siebert, *An Introduction to 3D Computer Vision Techniques and Algorithms*, 1st ed. Wiley-Blackwell, 2009.
- [3] F. Vahid and T. Givargis, “Embedded System Design: A Unified Hardware/Software Approach,” p. 103, 1999.
- [4] T. Bonny and J. Henkel, “Huffman-based code compression techniques for embedded processors,” *ACM Transactions on Design Automation of Electronic Systems*, vol. 15, no. 4, pp. 1–37, Sept. 2010.
- [5] M. Ros and P. Sutton, “Compiler optimization and ordering effects on VLIW code compression,” in *Proceedings of the international conference on Compilers, architectures and synthesis for embedded systems - CASES '03*. New York, New York, USA: ACM Press, 2003, p. 95.
- [6] K. Kissell, “MIPS16: High-density MIPS for the Embedded Market,” *Silicon Graphics MIPS Group*, pp. 559–571, 1997.
- [7] K. A. Publishers, *Codesign for real-time video applications*, Aug. 1998, vol. 36, no. 3.
- [8] D. Salomon and G. Motta, “Video Compression,” in *Handbook of Data Compression*. London: Springer London, 2010, pp. 855–952.

-
- [9] M. Piccardi, “Background subtraction techniques: a review,” in *IEEE International Conference on Systems, Man and Cybernetics*, 2004, pp. 3099–3104.
- [10] C. Savitha, M. ChidanandaMurthy, and M. Kurian, “Feature points to detect motion,” in *International Conference on Sustainable Energy and Intelligent Systems*, no. Seiscon. Iet, 2011, pp. 785–788.
- [11] S. S. Beauchemin and J. L. Barron, “The computation of optical flow,” *ACM Computing Surveys*, vol. 27, no. 3, pp. 433–466, Sept. 1995.
- [12] J. Weng, “A theory of image matching,” *Proceedings Third International Conference on Computer Vision*, no. d, pp. 0–9, 1990.
- [13] A. Waxman, J. Wu, and F. Bergholm, “Convected activation profiles and the measurement of visual motion,” *Proceedings CVPR '88: The Computer Society Conference on Computer Vision and Pattern Recognition*, no. 60, 1988.
- [14] J. Weber and J. Malik, “Robust computation of optical flow in a multi-scale differential framework,” *International Journal of Computer Vision*, vol. 14, no. 1, pp. 67–81, Jan. 1995.
- [15] C. Fermüller, D. Shulman, and Y. Aloimonos, “The Statistics of Optical Flow,” *Computer Vision and Image Understanding*, vol. 82, no. 1, pp. 1–32, Apr. 2001.
- [16] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, “Performance of optical flow techniques,” *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, Feb. 1992.
- [17] —, “Performance of optical flow techniques,” *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, Feb. 1994.
- [18] A. D. Fleet, David J. Jepson, *International Journal of Computer Vision*.

-
- [19] B. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision,” *Internal Joint Conference on Artificial Intelligence*, vol. 130, pp. 121–129, 1981.
- [20] B. Galvin, B. McCane, K. Novins, D. Mason, and S. Mills, “Recovering Motion Fields: An Evaluation of Eight Optical Flow Algorithms,” in *Proceedings of the British Machine Vision Conference 1998*. British Machine Vision Association, 1998, pp. 20.1–20.10.
- [21] B. McCane, B. Galvin, and K. Novins, “On the evaluation of optical flow algorithms,” Tech. Rep., 1998.
- [22] R. Walczyk, A. Armitage, and D. Binnie, “Comparative study on connected component labeling algorithms for embedded video processing systems.” in *IPCV'10*. Las Vegas, USA: CSREA Press, 2010.
- [23] R. M. Haralick, *Real-Time Parallel Computing Image Analysis*. Plenum Publishing Corporation, 1981.
- [24] F. Chang, C.-J. Chen, and C.-J. Lu, “A linear-time component-labeling algorithm using contour tracing technique,” *Computer Vision and Image Understanding*, vol. 93, no. 2, pp. 206–220, Feb. 2004.
- [25] B. P. Squires, “1. Introduction.” *Canadian Medical Association journal*, vol. 130, no. 5, p. 557, 1984.
- [26] K. Wu, E. Otoo, and K. Suzuki, “Optimizing two-pass connected-component labeling algorithms,” *Pattern Analysis and Applications*, vol. 12, no. 2, pp. 117–135, June 2009.
- [27] R. Haralick and Joo Hyonam, “2D-3D pose estimation,” in *9th International Conference on Pattern Recognition*, no. 1. IEEE Comput. Soc. Press, 1988, pp. 385–391.

-
- [28] C. H. Hansen, P. W. Alberti, D. L. Johnson, and N. Y. Samir, “Fundamentals of acoustics,” Geneva, p. NP, 2001.
- [29] R. E. Berg, “The Physics of Sound,” p. 953, 1982.
- [30] British Society of Audiology, “British Society of Audiology (BSA),” *Guidelines On, T H E Acoustics, O F Sound, Field Audiometry, Clinical Audiological Applications*, no. February, 2008.
- [31] K. S. Rao and Anil Kumar Vuppala, *Speech Processing in Mobile Environments*. Springer International Publishing, 2014.
- [32] B. Logan, “Mel-frequency cepstral coefficients for music modeling,” *Proceedings of the International Conference on Music Information Retrieval, Plymouth, Mass, USA*, 2000.
- [33] R. T. Edwards, “Time-Frequency Acoustic Processing and Recognition: Analysis and Analog VLSI Implementations,” Ph.D. dissertation, The Johns Hopkins University, Baltimore, Maryland, 1999.
- [34] O. M. Bouzid, G. Y. Tian, J. Neasham, and B. Sharif, “Investigation of sampling frequency requirements for acoustic source localisation using wireless sensor networks,” *Applied Acoustics*, vol. 74, no. 2, pp. 269–274, 2013.
- [35] O. M. Bouzid, G. Y. Tian, and A. other Other, “Envelope and Wavelet Transform for Sound Localisation at Low Sampling Rates in Wireless Sensor Networks,” *Journal of Sensors*, pp. 1–9, 2012.
- [36] C. H. Knapp and G. C. Carter, “The Generalized Correlation Method for Estimation of Time Delay,” *Ieee Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, 1976.
- [37] K. Sohraby, D. Minoli, and T. Znati, *Wireless Sensor Network Technology, Protocols, and Applications*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2007.

-
- [38] M. Wang, L. Ci, P. Zhang, and Y. Xu, "Acoustic Source Localization in Wireless Sensor Networks," *Workshop on Intelligent Information Technology Application.*, pp. 196–199, Dec. 2007.
- [39] F. Sivrikaya, "Time Synchronization in Sensor Networks : A Survey Computer Clocks and the Synchronization Problem," *IEEE Network*, vol. 18, no. 4, pp. 45–50, 2004.
- [40] Theodore S . Rappaport, *Wireless Communications : Principles and Practice*, second edi ed. Prentice Hall PTR, 2002.
- [41] M. Demirbas and S. Balachandran, "Robcast: A singlehop reliable broadcast protocol for wireless sensor networks," *Proceedings - International Conference on Distributed Computing Systems*, 2007.
- [42] N. A. Alrajeh, M. Bashir, and B. Shams, "Localization Techniques in Wireless Sensor Networks," *International journal of distributed*, 2013.
- [43] X. Sheng and Y. Hu, "Energy based acoustic source localization," *Information Processing in Sensor Networks*, vol. 2634, pp. 285–300, 2003.
- [44] K. Deng and Z. Liu, "Weighted Least-Squares Solutions for Energy-Based Collaborative Source Localization Using Acoustic Array," vol. 7, no. 1, pp. 159–165, 2007.
- [45] E. Xu, Z. Ding, and S. Dasgupta, "Source localization in wireless sensor networks from signal time-of-arrival measurements," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2887–2897, 2011.
- [46] M. abd elsalam mofeed Hassan Elkamchouchi, "Arrival (TDOA) Position Location Technique," in *National Radio Science Conference*, vol. 1, no. 3, 2005.
- [47] Kemmouche mohammed sadek, "Multisensor Multitarget Tracking Algorithms For Surveillance Systems," *Phd Thesis*, 2013.

-
- [48] C. Stauffer and W. Grimson, “Adaptive background mixture models for real-time tracking,” *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 246–252, 1999.
- [49] P. Dickinson and A. Hunter, “Scene modelling using an adaptive mixture of Gaussians in colour and space,” *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 64–69, 2005.
- [50] T. Yu, C. Zhang, M. Cohen, Y. Rui, and Y. Wu, “Monocular video foreground/background segmentation by tracking spatial-color Gaussian mixture models,” in *2007 IEEE Workshop on Motion and Video Computing, WMVC 2007*, 2007.
- [51] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.
- [52] T. Zhao and R. Nevatia, “Tracking multiple humans in complex situations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1208–1221, 2004.
- [53] D. Comaniciu, V. Ramesh, and P. Meer, “Real-time tracking of non-rigid objects using mean shift,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, no. 7, pp. 142–149, 2000.
- [54] B. Han, D. Comaniciu, and L. Davis, “Sequential kernel density approximation through mode propagation : Applications to background modelling,” in *Asian Conf. on Computer Vision*, 2004.
- [55] S. Rowe and A. Blake, “Statistical Background Modelling for Tracking with a Virtual Camera.” *Proceedings of the British Machine Vision Conference*, pp. 42.1–42.10, 1995.

-
- [56] F. Porikli and O. Tuzel, "Bayesian background modeling for foreground detection," *Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, p. 55, 2005.
- [57] D. Gutesst, M. TrajkoviCj, E. Cohen-Solalt, D. Lyons, and A. K. Jain, "A background model initialization algorithm for video surveillance," in *IEEE International Conference on Computer Vision*, 2001.
- [58] N. Friedman and S. Russell, "Image segmentation in video sequences: a probabilistic approach," in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, 1997.
- [59] E. Hayman and J.-o. Eklundh, "Statistical background subtraction for a mobile observer," *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 67–74 vol.1, 2003.
- [60] P. Kaewtrakulpong and R. Bowden, "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection," *Advanced Video Based Surveillance Systems*, pp. 1–5, 2001.
- [61] D. S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 827–832, May 2005.
- [62] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," *Proceedings of the 17th International Conference on Pattern Recognition*, pp. 28–31 Vol.2, 2004.
- [63] R. Klette, "Robust background subtraction and maintenance," *Proceedings of the 17th International Conference on Pattern Recognition*, pp. 90–93 Vol.2, 2004.
- [64] Y. Raja, S. J. McKenna, and S. Gong, "Segmentation and Tracking Using Colour Mixture Models," *Asian Conference on Computer Vision, volume 1351 of Lecture Notes in Computer Science*, pp. 607–614, 1998.

-
- [65] M. Xu and T. Ellis, "Illumination-Invariant Motion Detection Using Colour Mixture Models." *Bmvc*, pp. 163–172, 2001.
- [66] G. D. Finlayson, S. D. Hordley, J. A. Marchant, and C. M. Onyango, "Colour invariance at a pixel," in *11th British Machine Vision Conference*, 2000, pp. 13–22.
- [67] J. a. Marchant and C. M. Onyango, "Shadow-invariant classification for scenes illuminated by daylight." *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 17, no. 11, pp. 1952–1961, 2000.
- [68] O. Javed, K. Shafique, and M. Shah, "A hierarchical approach to robust background subtraction using color and gradient information," *Workshop on Motion and Video Computing, 2002. Proceedings.*, pp. 22–27.
- [69] L. Li and M. K. H. Leung, "Integrating intensity and texture differences for robust change detection," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 105–112, 2002.
- [70] L. Maddalena, A. Petrosino, and S. Member, "A Self-organizing approach to background subtraction for visual surveillance applications," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1–10, 2008.
- [71] S.-C. Huang and B.-H. Do, "Radial Basis Function Based Neural Network for Motion Detection in Dynamic Scenes." *IEEE transactions on cybernetics*, vol. 44, no. 1, pp. 114–125, 2013.
- [72] E. J. Palomo, E. Domínguez, R. M. Luque, and J. Muñoz, "Image hierarchical segmentation based on a GHSOM," *Lecture Notes in Computer Science*, vol. 5863 LNCS, pp. 743–750, 2009.
- [73] W. Kim and C. Kim, "Background subtraction for dynamic texture scenes using fuzzy color histograms," *IEEE Signal Processing Letters*, vol. 19, no. 3, pp. 127–130, 2012.

-
- [74] F. C. Cheng, S. C. Huang, and S. J. Ruan, "Scene analysis for object detection in advanced surveillance systems using laplacian distribution model," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 41, no. 5, pp. 589–598, 2011.
- [75] S.-C. Huang, "An Advanced Motion Detection Algorithm With Video Quality Analysis for Video Surveillance Systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 1–14, 2011.
- [76] J. M. Guo, C. H. Hsia, Y. F. Liu, M. H. Shih, C. H. Chang, and J. Y. Wu, "Fast background subtraction based on a multilayer codebook model for moving object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1809–1821, 2013.
- [77] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 597–610, 2013.
- [78] L. Xiong, X. Chen, and J. Schneider, "Direct robust matrix factorization for anomaly detection," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 844–853, 2011.
- [79] N. Wang, T. Yao, J. Wang, and D. Yeung, "A probabilistic approach to robust matrix factorization," *Computer Vision ECCV*, 2012.
- [80] M. S. Kemouche and N. Aouf, "A GMM approximation with merge and split for nonlinear non-Gaussian tracking," in *13th International Conference on Information Fusion*, 2010, pp. 1–6.
- [81] J. S. J. Shi and C. Tomasi, "Good features to track," *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pp. 593–600, 1994.

-
- [82] K. Nickels and S. Hutchinson, “Estimating uncertainty in SSD-based feature tracking,” *Image and Vision Computing*, vol. 20, no. 1, pp. 47–58, Jan. 2002.
- [83] a. Singh, “An estimation-theoretic framework for image-flow computation,” [1990] *Proceedings Third International Conference on Computer Vision*, pp. 168–177, 1990.
- [84] Y. Kanazawa and K. Kanatani, “Do we really have to consider covariance matrices for image features?” *Proceedings Eighth IEEE International Conference on Computer Vision*, vol. 2, pp. 301–306, 2001.
- [85] N. Goyette, “Changetection . net : A New Change Detection Benchmark Dataset,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–8.
- [86] H. Shih-Chia and C. Bo-Hao, “Highly accurate moving object detection in variable bit rate video-based traffic monitoring systems,” *IEEE Trans Neural Netw Learn Syst*, vol. 24, no. 12, pp. 1920–1931, 2013.
- [87] P. Chen, P. Ahammad, C. Boyer, S.-i. Huang, L. Lin, E. Lobaton, M. Meingast, S. Oh, S. Wang, P. Yan, A. Y. Yang, C. Yeo, L.-c. Chang, J. D. Tygar, S. S. Sastry, and I. Technology, “CITRIC : A LOW-BANDWIDTH WIRELESS CAMERA NETWORK PLATFORM,” in *ACM/IEEE Conference on Distributed Smart Cameras (ICDSC)*, 2008.
- [88] O. T. Inc, “OV9655 Color CMOS SXGA CAMERACHIP with OmniPixel Technology Datasheet,” Tech. Rep., 2006.
- [89] Intel Corp, “Intel PXA270 Processor Electrical, Mechanical and Thermal Specication Datasheet,” 2004.
- [90] Z. Yi and F. Liangzhong, “Moving object detection based on running average background and temporal difference,” *IEEE International*

- Conference on Intelligent Systems and Knowledge Engineering*, pp. 270–272, Nov. 2010.
- [91] T. P. Thompson, “Detecting moving objects,” *international journal of computer vision*, 1990.
- [92] Y. Z. Zhang, S. Kiselewich, W. Bauson, and R. Hammoud, “Robust Moving Object Detection at Distance in the Visible Spectrum and Beyond Using A Moving Camera,” *Conference on Computer Vision and Pattern Recognition Workshop*, 2006.
- [93] J. G. Ninad Thakoor and H. Chen, “Automatic Object Detection In Video Sequences With Camera In Motion,” *Proceedings of Advanced Concepts for Intelligent Vision Systems*, 2004.
- [94] R. Cucchiara, A. Prati, and R. Vezzani, “Advanced Video Surveillance with Pan Tilt Zoom Cameras,” *Proc of Workshop on Visual Surveillance VS at ECCV*, 2006.
- [95] S. W. Kim, K. Yun, K. M. Yi, S. J. Kim, and J. Y. Choi, “Detection of moving objects with a moving camera using non-panoramic background model,” *Machine Vision and Applications*, vol. 24, no. 5, pp. 1015–1028, Oct. 2012.
- [96] S. Kwak, T. Lim, W. Nam, B. Han, and J. H. Han, “Generalized background subtraction based on hybrid inference by belief propagation and Bayesian filtering,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2174–2181.
- [97] S. Shimizu, K. Yamamoto, C. Wang, Y. Sato, H. Tanahashi, and Y. Niwa, “Moving object detection with mobile stereo omni-directional system (SOS) based on motion compensatory inter-frame depth subtraction,” in *Proceedings - International Conference on Pattern Recognition*, vol. 3, 2004, pp. 248–251.

-
- [98] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-Up Robust Features (SURF),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [99] W. Niehsen, “Information fusion based on fast covariance intersection filtering,” in *Proceedings of the 5th International Conference on Information Fusion*, vol. 2, 2002, pp. 901–904.
- [100] S. A. Imran and N. Aouf, “Robust L homography estimation using reduced image feature covariances from an RGB image,” *Journal of Electronic Imaging*, vol. 21, no. 4, p. 043022, 2012.
- [101] W. Li and Y. Jia, “H infinity filtering for a class of nonlinear discrete time systems based on unscented transform,” *Signal Processing*, vol. 90, no. 12, pp. 3301–3307, 2010.
- [102] M. Boulekhour and N. Aouf, “L infinity norm based solution for visual odometry,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8048 LNCS, no. PART 2, 2013, pp. 185–192.
- [103] M. Kharbat and N. Aouf, “Dense optical flow via robust data fusion,” *Signal, Image and Video Processing*, vol. 5, no. 2, pp. 203–215, 2011.
- [104] G. Yu and G. Sapiro, “Solving inverse problems with piecewise linear estimators from Gaussian mixture models,” *IEEE transaction on image processing*, vol. 21, no. 5, pp. 2481–2499, 2012.
- [105] D. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999.
- [106] B. Zeisl, P. F. Georgel, F. Schweiger, E. Steinbach, and N. Navab, “Estimation of Location Uncertainty for Scale Invariant Feature Points,” *Proceedings of the British Machine Vision Conference*, pp. 57.1–57.12, 2009.

-
- [107] H. Zhou, H. Kong, J. M. Alvarez, D. Creighton, and S. Nahavandi, “Fast road detection and tracking in aerial videos,” in *IEEE Intelligent Vehicles Symposium, Proceedings*. Ieee, June 2014, pp. 712–718.
- [108] M. Majji, M. Diz, and D. Truong, “a Least Squares Solution for Estimation of a Planar Homography,” pp. 2333–2340.
- [109] Dan Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. John Wiley & Sons, Inc., 2006.
- [110] J.-y. Bouguet, “Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm,” *In Practice*, vol. 1, no. 2, pp. 1–9, Nov. 1999.
- [111] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*, 2008.
- [112] T. Olson and F. Brill, “Moving Object Detection and Event Recognition Algorithms for Smart Cameras,” in *Proc. DARPA Image Understanding Workshop*, 1997, pp. 159–175.
- [113] R. Collins, X. Zhou, and S. K. Teh, “An Open Source Tracking Testbed and Evaluation Web Site,” in *Pets*, 2005.
- [114] G. Padmavathi, “A Study on Vehicle Detection and Tracking Using Wireless Sensor Networks,” *Wireless Sensor Network*, vol. 02, no. 02, pp. 173–185, 2010.
- [115] T. Damarla, L. M. Kaplan, and G. T. Whipps, “Sniper localization using acoustic asynchronous sensors,” *IEEE Sensors Journal*, vol. 10, no. 9, pp. 1469–1478, 2010.
- [116] S. J. Cooke, J. D. Midwood, J. D. Thiem, P. Klimley, M. C. Lucas, E. B. Thorstad, J. Eiler, C. Holbrook, and B. C. Ebner, “Tracking animals

- in freshwater with electronic tags: past, present and future,” *Animal Biotelemetry*, vol. 1, no. 1, p. 5, 2013.
- [117] A. Shareef and Y. Zhu, “Localization Using Extended Kalman Filters in Wireless Sensor Networks,” in *Kalman Filter Recent Advances and Applications*, 2009, no. April, p. 24.
- [118] T. Ajdler, I. Kozintsev, R. Lienhart, and M. Vetterli, “Acoustic source localization in distributed sensor networks,” *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers.*, vol. 2, pp. 1328–1332, 2004.
- [119] K. D. A. Hashemi-Sakhtsari, “Recursive least squares solution to source tracking using time difference of arrival,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, no. 3, 2004, pp. 3–6.
- [120] Y. Weng, W. Xiao, and L. Xie, “Total Least Squares Method for Robust Source Localization in Sensor Networks Using TDOA Measurements,” pp. 1–8, 2011.
- [121] S. Gratton, A. S. Lawless, and N. K. Nichols, “Approximate Gauss-Newton methods for nonlinear least squares problems,” *SIAM Journal on Optimization*, vol. 18, no. 1, pp. 106–132, 2007.
- [122] C. V. L. GH Golub, “An analysis of the total least squares problem.pdf,” *SIAM Journal on Numerical Analysis*, 1980.
- [123] G. H. Golub, P. C. Hansen, and D. P. O’Leary, “Tikhonov Regularization and Total Least Squares,” *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 1, pp. 185–194, Jan. 1999.
- [124] G. H. Golub, “Some Modified Matrix Eigenvalue Problems,” pp. 318–334, 1973.

-
- [125] R. Adcock, "Note on the method of least squares." *Analyst*, vol. 4, no. 6, pp. 183,184, 1877.
- [126] P. Sprent, "Models in Regression and Related Topics," *London: Methuen*, vol. 21, no. 1, p. 91, Mar. 1969.
- [127] L. J. Gleser, "Estimation in a Multivariate "Errors in Variables" Regression Model: Large Sample Results," pp. 24–44, 1981.
- [128] J. E. D. J. Mei and H. H. W., "Two new unconstrained optimization algorithms which use function and gradient values," *Journal of Optimization Theory and Applications*, pp. 453–482, 1979.
- [129] J. R. Taylor, *Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements*,, 2nd ed. University Science Books, 1997, vol. 101.
- [130] S. Ghosh, *Distributed systems: an algorithmic approach*, second edi ed. Chapman, Publisher Crc, Hall, 2014.
- [131] M. I. A. Lourakis and A. A. Argyros, "Is Levenberg-Marquardt the Most Efficient Optimization Algorithm for Implementing Bundle Adjustment ?" in *ICCV '05 - 10th IEEE International Conference on Computer Vision*, 2005, pp. 1526 – 1531.
- [132] D. K. Chao-i Chen, Dusty Sargent, Chang-ming Tsai, Yuan-fang Wang, "Uniscale multi-view registration using double dog-leg method," *SPIE Medical Imaging*, vol. 7261, no. Ddl, 2009.
- [133] J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading." *Nature*, vol. 381, no. 6577, pp. 66–68, 1996.
- [134] W. H. I. P. Sumbly, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, p. 212, 1954.

-
- [135] Q. Summerfield, “Some preliminaries to a comprehensive account of audio-visual speech perception,” in *Hearing by Eye: The Psychology of Lip-reading*, 1987, pp. 3–51.
- [136] M. J. McGurk H, “Hearing lips and seeing voices,” *nature*, 1976.
- [137] G. Friedland, C. Yeo, and H. Hung, “Visual speaker localization aided by acoustic models,” *Proceedings of the 17th ACM international conference on Multimedia*, pp. 195–202, 2009.
- [138] Z. Li, T. Herfet, M. Grochulla, and T. Thormahlen, “Multiple active speaker localization based on audio-visual fusion in two stages,” *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 262–268, 2012.
- [139] R. Chellappa, G. Qian, and Q. Zheng, “Vehicle detection and tracking using acoustic and video sensors,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, no. 4, 2004, pp. 793–796.
- [140] V. Cevher, a.C. Sankaranarayanan, J. McClellan, and R. C. R. Chellappa, “Target Tracking Using a Joint Acoustic Video System,” *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 1–12, 2007.
- [141] S. Spors, R. Rabenstein, and N. Strobel, “A Multi-Sensor Object Localization System,” in *IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance*, 2003.
- [142] B. Bunin, A. Sutin, G. Kamberov, H.-S. Roh, B. Luczynski, and M. Burlick, “Fusion of acoustic measurements with video surveillance for estuarine threat detection,” *Proceedings of SPIE*, vol. 6945, pp. 694 514–694 514–12, 2008.
- [143] P. V. Hengel, “Verbal aggression detection in complex social environments,” *Advanced Video and Signal*, pp. 15–20, 2007.

-
- [144] S. Julier and J. Uhlmann, “A non-divergent estimation algorithm in the presence of unknown correlations,” *Proceedings of the American Control Conference*, vol. 4, pp. 2369–2373, 1997.
- [145] —, “General Decentralized Data Fusion With Covariance Intersection (CI),” in *Handbook of Multisensor Data Fusion*, 2001, no. Ci.
- [146] P. Levis and D. Gay, “TinyOS Programming,” *ReVision*, vol. 28, p. 206, 2009.
- [147] J. Elson, L. Girod, and D. Estrin, “Fine-grained network time synchronization using reference broadcasts,” in *the 5th symposium on Operating systems design and implementation*. Boston, 2002, pp. 147–163.
- [148] D. Gay, P. Levis, R. von Behren, M. Welsh, E. Brewer, and D. Culler, “The nesC language,” *ACM SIGPLAN Notices*, vol. 38, no. 5, p. 1, 2003.
- [149] M. a. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” pp. 381–395, 1981.
- [150] M. Li, Peihua Xianzhe, “Robust Acoustic Source Localization with TDOA Based Ransac Algorithm,” in *International Conference on Intelligent Computing*, 2009, pp. 222–227.
- [151] B. K. Horn, “Tsais Camera Calibration Method Revisited,” Department of Electrical Engineering and Computer Science Massachusetts Institute of Technology, Tech. Rep., 2000.
- [152] C. B. H. Durrant-white, B. Rao, and B. Steer, “Centralized and decentralized kalman filter technique for tracking, navigation, and control,” University of Rochester, Computer science, Tech. Rep., 1989.

-
- [153] Y. Bar-Shalom and H. C. H. Chen, "Multisensor track-to-track association for tracks with dependent errors," *43rd IEEE Conference on Decision and Control*, vol. 3, no. 1, pp. 3–14, 2004.
- [154] Y. Bar-shalom and X.-r. Li, *Multitarget-Multisensor Tracking : Application and Advances*, Yaakov Bar-Shalom William Dale Blair, Ed. Artech House Publishers, 2000.
- [155] N. G. Wah and Y. Rong, "Comparison of decentralized tracking algorithms," in *Proceedings of the 6th International Conference on Information Fusion*, vol. 1, 2003, pp. 107–113.
- [156] Z. Liu, "An Evaluation of Several Fusion Algorithms for Multi-sensor Tracking System," *Journal of Information & computational Science*, vol. 10, no. October, pp. 2101–2109, 2010.
- [157] I. Liggins, M.E., C.-Y. C. C.-Y. Chong, I. Kadar, M. Alford, V. Vannicola, and S. Thomopoulos, "Distributed fusion architectures and algorithms for target tracking," *Proceedings of the IEEE*, vol. 85, no. 1, 1997.
- [158] B. Rao and H. Durrant-Whyte, "Fully decentralised algorithm for multi-sensor Kalman filtering," p. 413, 1991.
- [159] D. Papageorgiou and M. Holender, "Track-to-track association and ambiguity management in the presence of sensor bias," *12th International Conference on Information Fusion*, vol. 6, no. 2, 2009.
- [160] E. Wan and R. V. D. Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proceedings of the Adaptive Systems for Signal Processing, Communications, and Control Symposium*, 2000, pp. 153–158.
- [161] X. Li, "Visual Navigation in Unmanned Air Vehicles with Simultaneous Location and Mapping," Ph.D. dissertation, Cranfield University, 2014.

- [162] J.-P. Renno, P. Remagnino, and G. Jones, “Learning surveillance tracking models for the self-calibrated ground plane,” *Multisensor surveillance systems*, 2003.